Utilisation des modèles de co-clustering pour l'analyse exploratoire des données

Romain Guigourès

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ PARIS 1 PANTHÉON-SORBONNE

ÉCOLE DOCTORALE : Sciences Mathématiques de Paris Centre (ED 386)

Par Romain GUIGOURÈS

POUR OBTENIR LE GRADE DE
DOCTEUR

SPÉCIALITÉ : Mathématiques Appliquées

Utilisation des modèles de co-clustering pour l'analyse exploratoire des données

Directeur de thèse : Fabrice ROSSI
Co-encadrant : Marc BOULLÉ

Soutenue le : 4 Décembre 2013

Devant la commission d'examen formée de :

M. Mohamed NADIF	Professeur	LIPADE – Université Paris Descartes	Rapporteur
M. Gilbert SAPORTA	Professeur	Conservatoire National des Arts et Métiers	Rapporteur
M. Emmanuel VIENNET	Professeur	L2TI – Université Paris 13	Président
M. Gilles BISSON	Docteur	LIG – Université de Grenoble	Examinateur
M. Vincent BLONDEL	Professeur	Université Catholique de Louvain	Examinateur
M. Fabrice ROSSI	Professeur	SAMM – Université Paris 1 Panthéon-Sorbonne	Directeur
M. Marc BOULLÉ	Docteur	Orange Labs Lannion	Co-encadrant

Remerciements

Je tiens tout dabord à remercier Monsieur Marc Boullé et Monsieur Fabrice Rossi d'avoir accepté d'encadrer cette thèse. Je les remercie pour leur disponibilité, leurs précieux conseils et leur implication durant ces trois années de thèse. Leurs encouragements et leur engagement m'ont permis d'enrichir mes connaissances, d'acquérir des méthodes de travail et de mener à bien mes travaux.

Je souhaite également remercier Monsieur Gilbert Saporta et Monsieur Mohamed Nadif de m'avoir fait l'honneur d'accepter d'être rapporteurs de ma thèse. Je remercie Monsieur Gilles Bisson, Monsieur Vincent Blondel et Monsieur Emmanuel Viennet pour avoir participé à l'évaluation de mes travaux.

Ce travail a été mené au sein de l'équipe Profiling & Datamining d'Orange Labs Lannion où j'ai pu bénéficier d'un cadre de travail exceptionnel. Je tiens à remercier tous les membres de l'équipe pour leurs échanges et la sympathie qu'ils m'ont témoignée. Je remercie toutes les personnes qui m'ont aidé à mener à bien cette thèse et qui ont consacré du temps à la relecture du manuscrit. Un remerciement particulier à Monsieur Fabrice Clérot pour les nombreuses discussions qui m'ont permis d'enrichir mes connaissances scientifiques et ma culture générale. Je remercie également Monsieur Dominique Gay pour son aide et ses encouragements qui ont su me motiver.

Je remercie ma famille et mes amis pour le soutien qu'ils m'ont apporté pendant ces trois années de thèse.

Résumé

Le co-clustering est une technique de classification consistant à réaliser une partition simultanée des lignes et des colonnes d'une matrice de données. Parmi les approches existantes, MODL permet de traiter des données volumineuses et de réaliser une partition de plusieurs variables, continues ou nominales. Nous utilisons cette approche comme référence dans l'ensemble des travaux de la thèse et montrons la diversité des problèmes de data mining pouvant être traités, comme le partitionnement de graphes, de graphes temporels ou encore le clustering de courbes.

L'approche MODL permet d'obtenir des résultats fins sur des données volumineuses, ce qui les rend difficilement interprétables. Des outils d'analyse exploratoire sont alors nécessaires pour les exploiter. Afin de guider lutilisateur dans linterprétation de tels résultats, nous définissons plusieurs outils consistant à simplifier des résultats fins afin d'en avoir une interprétation globale, à détecter les clusters remarquables, à déterminer les valeurs représentatives de leurs clusters et enfin à visualiser les résultats. Les comportements asymptotiques de ces outils d'analyse exploratoire sont étudiés afin de faire le lien avec les approches existantes.

Enfin une application sur des comptes-rendus d'appels de l'opérateur Orange, collectés en Côte d'Ivoire, montre l'intérêt de l'approche et des outils d'analyse exploratoire dans un contexte industriel.

Abstract

Co-clustering is a clustering technique aiming at simultaneously partitioning the rows and the columns of a data matrix. Among the existing approaches, MODL is suitable for processing huge data sets with several continuous or categorical variables. We use it as the baseline approach in this thesis. We discuss the reliability of applying such an approach on data mining problems like graphs partitioning, temporal graphs segmentation or curve clustering.

MODL tracks very fine patterns in huge data sets, that makes the results difficult to study. That is why, exploratory analysis tools must be defined in order to explore them. In order to help the user in interpreting the results, we define exploratory analysis tools aiming at simplifying the results in order to make possible an overall interpretation, tracking the most interesting patterns, determining the most representative values of the clusters and visualizing the results. We investigate the asymptotic behavior of these exploratory analysis tools in order to make the connection with the existing approaches.

Finally, we highlight the value of MODL and the exploratory analysis tools owing to an application on call detailed records from the telecom operator Orange, collected in Ivory Coast.

Table des matières

1

Introduction

Depuis quelques années, on assiste à une augmentation significative du volume des données. Le domaine des télécommunications est particulièrement touché par cette croissance : le taux de pénétration de la téléphonie mobile en France est passé de 64% à 114% en dix ans et le trafic mensuel de SMS a été multiplié par cent dans la même période[1]. Afin de mieux comprendre les usages dans une optique de développement des services, les entreprises s'intéressent à l'extraction de l'information présente dans leurs données. Ce processus d'extraction et de restitution est appelé *fouille de données* (ou *data mining*).

Les techniques de data mining se sont adaptées à la croissance du volume des données. L'amélioration des performances des algorithmes et l'augmentation des moyens de calcul permettent aujourd'hui de traiter d'importantes quantités de données. Mais ces avancées sont à l'origine de nouvelles difficultés. Plus les données sont nombreuses, plus la quantité d'information fiable potentiellement extraite est importante. Il existe de nombreuses approches de data mining permettant une restitution fine de l'information extraite des données. Cependant, la complexité des résultats produits peut être un réel obstacle pour l'interprétation. Il est alors nécessaire que l'utilisateur soit guidé afin d'exploiter les résultats obtenus. Dans ce cas, la mise en place d'outils d'analyse exploratoire s'avère nécessaire. Il faut permettre à l'utilisateur d'avoir une vue des résultats aussi bien macroscopique – pour comprendre la structure générale des données – que microscopique, pour détecter les phénomènes de niches, c'est-à-dire des segments dans les données présentant un fort intérêt opérationnel mais contenant peu d'individus.

Le data mining peut être divisé en deux grandes sous-familles. Dans les approches supervisées, on cherche à prédire la valeur d'une variable cible continue (régression) ou nominale (classification supervisée), à partir des attributs descriptifs des données. Dans le domaine des télécommunications, on utilise notamment ce type d'études pour attribuer un score à un client en fonction de ses caractéristiques. On peut par exemple chercher à prédire l'attrition (perte de clientèle) ou l'appétence du client (volonté de souscrire à un service). L'autre grande famille regroupe les approches non-supervisées, parmi lesquelles on peut

1. Observatoire des marchés des communications électroniques en France, actes de l'ARCEP, Juillet 2013.

citer l'estimation de densité, la détection de motifs et la classification, que nous désignons par le terme *clustering* dans la suite du manuscrit. Le clustering consiste à grouper des individus similaires. Par exemple, dans les problèmes de segmentation marketing, on crée des groupes d'individus ayant le même profil en fonction d'attributs, comme l'âge, la catégorie socio-professionnelle, les habitudes de consommation, etc.

Mais les approches de clustering classique ne sont pas adaptées à tous les problèmes non-supervisés. Dans certaines études, on cherche à effectuer des regroupement d'objets plus complexes. Parmi elles, le clustering de nœuds dans les graphes a connu un nouvel engouement ces dernières années, lié notamment à l'émergence des réseaux sociaux. L'évolution temporelle de ces graphes et du rôle des acteurs des réseaux est un problème qui suscite de plus en plus l'intérêt des industriels. Dans le domaine des télécommunications, ce type d'études est utilisé afin de comprendre les profils socio-économiques des usagers du réseau. Comprendre les habitudes comportementales de la population d'un quartier, d'une ville ou d'une région est important dans le développement de l'activité d'un opérateur de téléphonie mobile.

Afin de répondre à cette demande d'analyse de données plus complexes, nous utilisons le *co-clustering*. Cette technique consiste à réaliser un clustering simultané des lignes et des colonnes d'une matrice de données. Les dimensions de cette matrice peuvent être des individus et des attributs (Hartigan, 1972), ou alors des variables dont on cherche à faire des clusters de modalités (Dhillon *et al.*, 2003). Les applications d'une telle approche sont nombreuses.

De la même manière que les approches de clustering, les approches de co-clustering sont conditionnées par le choix de plusieurs paramètres comme le nombre de clusters, la mesure de dissimilarité ou encore la représentation des données. L'approche MODL (Boullé, 2007) ne requérant pas de paramètre de l'utilisateur, nous avons fait le choix d'axer cette thèse autour de cette approche de co-clustering.

Avec l'augmentation du trafic et du nombre d'usagers, ces segmentations sont de plus en plus complexes à interpréter et nécessitent la mise en place d'une méthodologie d'analyse exploitant au mieux les résultats. Pour cela, nous proposons, dans cette thèse, un ensemble d'outils d'analyse exploratoire déduits de l'approche de co-clustering utilisée pour mener les analyses.

Les pays émergents connaissent actuellement une très forte croissance de la consommation de services téléphoniques. Il est donc primordial de comprendre les données dont nous disposons pour développer au mieux les usages et proposer des services adaptés. Dans cette thèse, nous menons une étude sur des comptes-rendus d'appels enregistrés en Côte-d'Ivoire en 2012. Le volume des communications étudiées nous permet d'avoir des résultats très fins. Notre contribution est de proposer une méthodologie et des outils d'analyse, pour permettre à l'utilisateur de se focaliser sur les observations les plus remarquables dans les résultats. Une fois l'information intéressante extraite, les résultats sont interprétés et validés par des experts du domaine.

La thèse se compose de cinq chapitres.

Le chapitre 2 est un état de l'art divisé en deux parties correspondant aux deux thématiques principales de la thèse. Dans un premier temps, nous nous intéressons aux difficultés liées au clustering. Nous nous intéressons également aux problèmes de l'interprétation et de l'exploration des résultats des algorithmes de clustering. La deuxième partie du chapitre est un état de l'art sur le co-clustering où plusieurs approches sont présentées, y compris l'approche MODL (Boullé, 2007), dont le fonctionnement est détaillé. Cette approche sert de support pour les chapitres suivants de la thèse.

Le chapitre 3 est une présentation de différents problèmes pouvant être traités par les approches de co-clustering. Nous montrons que le problème du clustering de nœuds dans les graphes peut être traité par un co-clustering en deux dimensions (ou biclustering). Nous proposons ensuite de traiter le problème de la segmentation des graphes temporels par application d'un co-clustering en trois dimensions (ou triclustering) et montrons en quoi cette formalisation est adaptée à ce type de problèmes. Nous traitons également le cas du clustering de courbes en considérant les données comme un ensemble de points de mesures décrits par l'identifiant de leur courbe et leurs valeurs en abscisse et ordonnée. On utilise dans ce cas un triclustering également. Enfin, nous proposons une extension au cas à d dimensions et présentons un ensemble de problèmes pouvant être traités par cette approche.

Le chapitre 4 est dédié à l'exploration et à l'exploitation des résultats du co-clustering. L'approche MODL permet d'obtenir des résultats optimaux très fins et informatifs. Dans ce chapitre, nous proposons deux axes d'analyse des résultats. Le premier consiste à détecter les clusters les plus atypiques. Nous définissons pour cela une mesure d'intérêt des clusters qui nous permet de détecter les niches au sein d'une partition fine. Le second consiste à détecter les individus les plus caractéristiques de leur cluster. Nous utilisons pour cela une mesure de typicité qui nous servira à étiqueter les clusters. Nous étudierons également, dans ce chapitre, une mesure de dissimilarité directement dérivée du critère optimisé par l'approche MODL. Cette mesure nous servira à simplifier les résultats trop complexes à l'aide d'une classification hiérarchique ascendante. Enfin, nous proposons des outils de visualisation basés sur des concepts de théorie de l'information, de manière à faciliter l'interprétation de la structure de co-clustering des données.

Le chapitre 5 est une application du co-clustering sur un problème lié à l'activité de l'opérateur Orange : l'analyse de compte-rendus d'appels collectés en Côte d'Ivoire. Cette application a pour but d'illustrer les différents modèles de co-clustering qui peuvent être utilisés pour traiter divers problèmes opérationnels. Nous proposons dans un premier temps une méthodologie d'analyse des données permettant d'avoir une seule démarche d'étude pour toutes les analyses menées sur ces données. Cette méthodologie permet de mieux comprendre l'utilisation des outils d'analyse précédemment introduits. La première analyse est une segmentation du territoire Ivoirien, menée en étudiant le graphe de communications entre les antennes du pays. Cette étude est approfondie en

ajoutant des composantes temporelles comme la date, le jour de la semaine ou encore l'heure de la journée, dans le but de caractériser les usages par zone géographique. Une seconde analyse consiste en l'étude des communications internationales au niveau des antennes, auxquelles on ajoute également des informations temporelles, ainsi que le type de communications afin de comprendre la nature des échanges avec l'international. Enfin, dans une dernière étude, on réalise une segmentation d'un échantillon d'utilisateurs en fonction de leur utilisation du réseau : position géographique, jour de la semaine, heure. Ainsi on mène une étude de la mobilité des utilisateurs.

Pour finir, une conclusion dresse le bilan des trois années de thèse, des travaux réalisés et des travaux futurs. Nous rappelons les différentes notions introduites dans la thèse, ainsi que les résultats obtenus. Nous positionnons également les travaux réalisés par rapport aux besoins opérationnels d'Orange en termes d'analyse exploratoire, en détaillant les éléments introduits dans la thèse qui sont aujourd'hui utilisés et en proposant d'étendre certains concepts faisant l'objet d'utilisation dans des études en cours.

2

Clustering et co-clustering

Dans le domaine de l'analyse de données, il existe deux grandes catégories d'approches. Les approches supervisées traitent les données étiquetées, et apprennent, à partir de données explicatives, à prédire une variable à expliquer. Au contraire, les approches non-supervisées n'ont pas de variables à expliquer mais cherchent la structure sous-jacente des données. Une de ces approches, la classification (ou *clustering*), a pour but de grouper des observations similaires. Les observations peuvent appartenir à un ou plusieurs groupes (ou *clusters*). Dans cette thèse, nous nous limitons au cas où les observations appartiennent à un seul cluster. Plus complexe que le clustering simple, le co-clustering est également une approche non-supervisée qui réalise un partitionnement simultané des dimensions d'une matrice de données.

Nous nous intéressons dans un premier temps à définir les concepts de base du clustering (choix des paramètres, de la mesure de dissimilarité...) et dans un second temps, nous introduisons la notion de co-clustering et présentons différentes approches dans un état de l'art.

2.1 Le clustering, définitions et concepts

Dans les problèmes de clustering, les données sont composées d'*observations* sans étiquette (ou *classe*), chacune décrite par plusieurs *variables*. On note $X = \{X_1, X_2, ..., X_d\}$ l'ensemble des d variables décrivant les *données* $\mathcal{D}$, un ensemble de m observations. Chaque observation est un vecteur u de dimension d tel que $\mathcal{D} = \{u^{(i)}\}_{i=1..m}$. Les données peuvent donc être représentées par une matrice de taille $m \times d$.

Il existe un très grand nombre d'algorithmes abordant le problème du clustering sous différents aspects. Ces différentes approches ne sont pas l'objet de la thèse et ne sont donc pas étudiées. Cependant, nous introduisons, dans ce chapitre, les notions clés des approches de clustering, afin de cibler les besoins en termes d'analyse exploratoire développée dans la thèse. Cet état de l'art sur le clustering est, de ce fait, succinct. Jain et Dubes (1988) proposent une étude détaillée des différents points abordés dans cette section.

2.1.1 Représentation des données

La représentation des données est une étape indispensable en amont du clustering. Bien souvent, une mauvaise représentation produit un clustering complexe et difficilement exploitable. Il se peut ainsi qu'un même algorithme produise un clustering satisfaisant et un clustering aberrant sur les mêmes données mais représentées différemment.

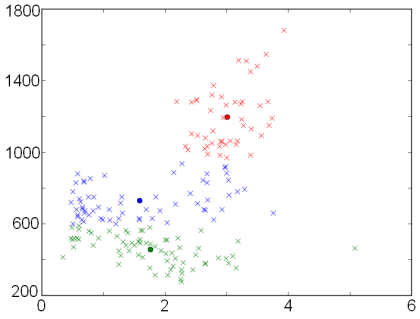
En fonction de la nature des variables traitées, leur représentation est différente. Nous considérons deux principaux types de variables :

- *les variables nominales* (ou *discrètes*, ou *catégorielles*) : variables prenant un nombre fini de valeurs non-ordonnées. Par exemple : un nom, une couleur, etc ;
- *les variables continues* (ou *numériques*) : variables prenant leurs valeurs dans $\mathbb{R}$ ou dans un de ses sous-ensembles. Par exemple : la taille, une distance, etc.

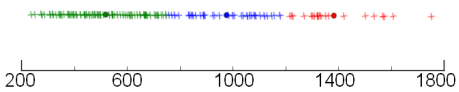
Les variables continues Bien que les données décrites par des variables continues soient directement exploitables, il est souvent préférable de les prétraiter afin d'exploiter au mieux l'information qu'elles apportent. En effet, admettons que nous ayons des variables avec des échelles de valeurs différentes. Si on applique un algorithme de clustering, basé sur un calcul de distance (voir section 2.1.2), les variables ayant les valeurs les plus élevées dominent les variables ayant des valeurs moindres et l'algorithme n'exploite pas l'information apportée par certaines variables. C'est ce qui est illustré par la figure 2.1.

Une solution possible à ce problème est de normaliser les données. Centrer et réduire les variables est une normalisation permettant de fixer la moyenne et la variance des valeurs des variables respectivement à 0 et à 1. Si cette normalisation présente l'avantage de préserver la dispersion des données et de ramener toutes les données à une même échelle, elle présuppose que les données

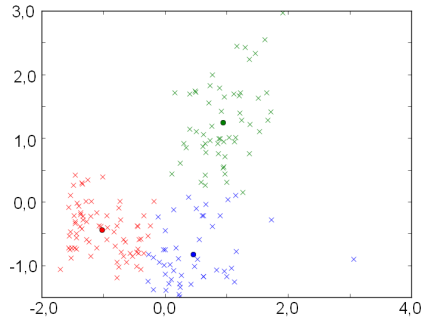
sont distribuées selon une loi proche de la loi normale et ignore la présence de données aberrantes (ou *outliers*), ce qui représente des hypothèses fortes, pas toujours vérifiées empiriquement dans les applications réelles. Une autre normalisation fréquemment utilisée, et présentant des propriétés similaires est la normalisation MinMax, elle consiste à projeter les observations sur l'intervalle $[0, 1]$.



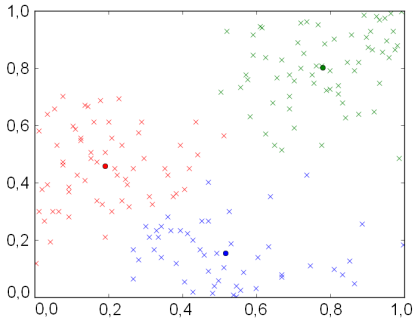
(a) Clustering sur des données non normalisées



(b) Clustering sur une seule des deux variables



(c) Clustering sur des données après centrage et réduction des variables



(d) Clustering sur des données après normalisation par statistique d'ordre

FIGURE 2.1 – Un clustering a été réalisé sur les données Wine (UCI). Seules deux variables avec des échelles différentes ont été conservées. L'algorithme des k-means a été appliqué, dans un premier temps sur des données non normalisées (a). On réalise que la variable ayant les valeurs les plus importantes a dominé l'autre en réalisant une projection des clusters sur cette même variable (b). Une normalisation des variable par centrage et réduction est effectuée puis l'algorithme k-means est appliqué (c). La même opération est ensuite réalisée avec une normalisation de type statistique d'ordre (d).

La normalisation par *statistique d'ordre* consiste à remplacer les valeurs des variables par leur rang, projeté sur $[0, 1]$. Chaque observation a donc un rang unique selon chacune des d variables. Cette normalisation a pour caractéristiques de supprimer la dispersion des données, au sens où on se ramène à une loi

uniforme. Ainsi l'impact des valeurs aberrantes est limité. De plus, les données deviennent invariantes par transformation monotone. L'impact sur le clustering de la normalisation est illustré par la figure 2.1.

Les variables nominales. Un clustering de données décrites par des variables nominales est complexe. Dans le cas continu, il est facile de quantifier la différence de valeurs prise par deux observations. En effet, admettons que l'on veuille faire un clustering de données où les observations sont des consommateurs décrits par une variable continue indiquant leur âge et une variable nominale indiquant leur parfum préféré de crème glacée. Il semble évident qu'un consommateur de 10 ans est plus proche d'un consommateur de 12 ans que de 50 ans, sur l'axe des âges. Cependant, il est difficile de dire si un consommateur préférant les glaces au café est plus proche du consommateur préférant les glaces au chocolat ou à la vanille. Il est courant d'appliquer une *recodage disjonctif complet* sur les données nominales. Il s'agit de remplacer la variable qualitative par des variables booléennes représentant les *valeurs* (ou *modalités*) de la variable remplacée. Les variables peuvent ainsi être traitées comme des variables numériques (Ralambondrainy, 1995). Mais cette transformation ne résout en rien le problème de la dissimilarité dans le cas nominal, il s'agit simplement de fixer la plupart du temps une dissimilarité de 1 entre toutes les modalités des variables. Le problème d'un tel pré-traitement des données est l'augmentation du nombre de variables, et donc de la complexité algorithmique. Cette transformation des variables nominale n'est pas nécessaire, nous verrons dans la section 4.3.1 qu'il existe des mesures de dissimilarités adaptées aux données nominales.

La sélection de variables Il est souvent préférable dans un clustering de ne garder que les variables les plus descriptives et les plus discriminantes. Dans le but d'avoir un clustering de meilleure qualité et plus interprétable (voir section 2.1.6), on cherche à éliminer les variables qui sont inintéressantes dans l'analyse. La première étape consiste à éliminer les variables de manière subjective, c'est-à-dire qu'elles sont enlevées des données à partir de connaissances a priori des données, en fonction des besoins de l'analyse. Dans un second temps, on peut observer des redondances dans les données qui apparaissent dans des variables différentes. Par exemple, si on souhaite faire un clustering sur une base de données de produits destinés à la vente, le prix peut être une variable très discriminante. Cependant, il est inutile de conserver le prix hors taxe et le prix avec les taxes dans l'analyse car ces deux variables sont très fortement corrélées. On cherche donc à éliminer ces variables afin d'avoir un clustering plus parcimonieux.

La sélection de variables fait l'objet d'une littérature abondante dans le cas des analyses supervisées, Guyon et Elisseeff (2003) ont réalisé un état de l'art des méthodes de sélection de variables dans ce contexte. Dans le cas non supervisé, la sélection de variables a été beaucoup moins étudiée. Des

techniques de réduction de la dimensionnalité ou d'extraction des variables ont été proposées, comme par exemple l'*Analyse en Composantes Principales* (ACP). Cette technique permet de transformer un ensemble de variables en un ensemble réduit de nouvelles variables, combinaisons linéaires des variables initiales. Les variables associées au coefficient le plus fort sont les plus informatives pour le clustering. Une autre technique consiste à filtrer les variables utiles à l'analyse. D'abord introduits dans le cadre de l'analyse supervisée, les approches de type *filter* (Guyon et Elisseeff, 2003) permettent de sélectionner un sous-ensemble de variables ou de classer ces variables en fonction d'un critère de qualité. Dash *et al.* (2002) proposent une adaptation au cas non supervisé de ces approches en se basant sur une mesure d'entropie. Une autre technique de sélection de variables populaire en classification supervisée est l'approche *wrapper* (Guyon et Elisseeff, 2003). Elle consiste à utiliser un prédicteur pour attribuer un score aux variables. Le prédicteur est entraîné sur plusieurs sous-ensembles de variables différents : celui qui maximise les performances est sélectionné. Une adaptation à la classification non-supervisée a été proposée par Dy et Brodley (2004). Une approche de clustering de type *Espérance-Maximisation* est appliquée et ses performances sont mesurées grâce à un critère de séparabilité des clusters ainsi que par maximum de vraisemblance.

D'autres approches de sélection de variables adaptées aux approches non-supervisées ont été étudiées en détails par Liu et Yu (2005).

2.1.2 La notion de dissimilarité

L'objectif du clustering est de grouper des observations similaires. Il est donc important de définir une notion de *similarité* (ou de *dissimilarité*). Les algorithmes de clustering optimisent un critère dérivé de cette mesure afin de déterminer la meilleure partition des données. Le résultat est donc très lié au choix de la dissimilarité entre observations. La mesure de dissimilarité peut être soit calculée sur les variables descriptives des observations, soit donnée directement. Ici, on se limite à l'étude des dissimilarités calculées à l'aide des variables.

Les distances La dissimilarité, entre deux observations est souvent définie comme une distance. Pour rappel, une distance est caractérisée par les propriétés suivantes :

- une distance est positive ou nulle : $D(u^{(i)}, u^{(j)}) \geq 0$,
- une distance est nulle si et seulement si les deux observations comparées sont identiques : $D(u^{(i)}, u^{(j)}) = 0 \Longleftrightarrow u^{(i)} = u^{(j)}$,
- une distance est symétrique : $D(u^{(i)}, u^{(j)}) = D(u^{(j)}, u^{(i)})$,
- une distance respecte l'inégalité triangulaire : $D(u^{(i)}, u^{(k)}) \leq D(u^{(i)}, u^{(j)}) + D(u^{(j)}, u^{(k)})$.

Définition 1. *Parmi les mesures de similarité utilisées dans le cadre de variables continues, les distances de Minkowski sont couramment utilisées. Elles*

sont définies de la manière suivante :

$$D_p(u^{(i)}, u^{(j)}) = \left(\sum_{k=1}^{d} |u_k^{(i)} - u_k^{(j)}|^p \right)^{\frac{1}{p}} \tag{2.1}$$

Notons que dans le cas où $p = 1$ on obtient la *distance de Manhattan*, qui correspond à la somme des valeurs absolues des différences entre les valeurs prises par deux observations sur chaque dimension. Pour $p = 2$, on a la *distance euclidienne*, qui est souvent utilisée pour évaluer la proximité des observations les unes par rapport aux autres. Cette distance est une mesure de similarité adaptée aux données avec des observations formant des clusters compacts et isolés et lorsque les variables sont indépendantes. Dans le cas contraire, s'il existe une corrélation entre les variables, alors cela reviendrait à donner plus de poids à certaines variables. Afin de corriger ce problème il est possible d'utiliser la *distance de Mahalanobis* qui prend en compte la covariance des variables (Duda *et al.*, 2001). Enfin, avec $p = \infty$, on obtient la distance de Tchebychev, qui correspond à la différence maximale entre les valeurs prises par deux observations sur chaque dimension.

Dans le cas où les données ne sont pas continues, la *distance de Hamming* permet de quantifier la différence entre deux ensembles de valeurs nominales en comptant le nombre de valeurs qui diffèrent d'un ensemble à l'autre. Ainsi la distance minimale entre deux observations est 0 si elles prennent les mêmes valeurs sur chacune des variables nominales. À l'inverse, la distance maximale est égale à $2 \times d$, c'est-à-dire deux fois le nombre de variables.

Définition 2. *La distance de Hamming D_H est définie de la manière suivante (Steane, 1996) :*

$$D_H(u^{(i)}, u^{(j)}) = \sum_{k=1}^{d} (u_k^{(i)} \oplus u_k^{(j)}) \tag{2.2}$$

où $\oplus$ est l'opérateur ou exclusif.

Une autre distance très utilisée est la *distance de Jaccard* (Duda *et al.*, 2001), qui correspond au ratio du nombre de valeurs communes prises par deux observations sur le nombre de valeurs différentes observées sur les deux observations, soustrait à 1. Ainsi, une distance de Jaccard est nulle si les deux observations sont identiques et vaut 1 si elles sont totalement différentes.

Définition 3. *La distance de Jaccard D_J est définie de la manière suivante (Duda* et al.*, 2001) :*

$$D_J(u^{(i)}, u^{(j)}) = 1 - \frac{|u^{(i)} \cap u^{(j)}|}{|u^{(i)} \cup u^{(j)}|} \tag{2.3}$$

De nombreuses mesures de similarités pour les variables nominales ont été définies et sont détaillées par Boriah *et al.* (2008).

D'après Wilson et Martinez (1997), lorsque les données sont décrites par des variables continues et nominales, il est possible de définir une distance euclidienne adaptée. Cette métrique est nommée HEOM (Heterogeneous Euclidean-Overlap Metric) et suggère de pré-traiter les données nominales en réalisant un codage disjonctif complet, de normaliser les données numériques par une statistique d'ordre, et de prendre la distance Euclidienne comme mesure de similarité sur les données ainsi préparées.

Dissimilarité entre clusters Dans certains problèmes, comme la classification hiérarchique ascendante, on peut être amené à calculer la dissimilarité entre clusters. Dans ce type d'approches de clustering, le clustering est initialisé avec autant de clusters que d'observations. À chaque étape, les deux clusters les plus similaires sont fusionnés jusqu'à ce qu'il n'y ait plus qu'un seul cluster. Ainsi on obtient une hiérarchie des clusters. Plusieurs mesures de similarités peuvent être utilisées, basées notamment sur la distance euclidienne. Cependant, la notion de dissimilarité entre deux clusters peut être définie de plusieurs façon, en fonction du *lien* choisi :

- lien minimal : la dissimilarité entre deux clusters est la distance minimale entre les observations de chacun des deux clusters,
- lien maximal : la dissimilarité entre deux clusters est la distance maximale entre les observations de chacun des deux clusters,
- lien moyen : la dissimilarité entre deux clusters est la distance moyenne entre toutes les observations de chacun des deux clusters,
- lien de Ward : la dissimilarité entre deux clusters est la distance entre les barycentres des clusters, pondérée par la moyenne harmonique des effectifs des clusters. Ce lien ne peut être utilisé que lorsque les données sont décrites par des variables continues.

Peu importe le lien choisi, la dissimilarité entre clusters n'est pas une distance. En effet, aucune d'elles ne respecte la propriété d'identité des indiscernables. Deux clusters différents peuvent avoir une dissimilarité nulle si le lien minimal ou le lien de Ward est utilisé. Plus formellement, soit c_i et c_j deux clusters et Δ la dissimilarité choisie, $c_i = c_j \Rightarrow \Delta(c_i, c_j) = 0$ mais la réciproque est fausse. D'autre part, les dissimilarités associées aux liens de Ward et minimal ne respectent pas l'inégalité triangulaire.
Au contraire, si le lien maximal ou le lien moyen est utilisé, une dissimilarité nulle n'est observée que si les observations des deux clusters sont toutes exactement identiques. Dans le cas contraire, si le lien maximal est utilisé, la dissimilarité entre deux clusters identiques est égale à leur diamètre et si le lien moyen est utilisé, elle correspond à la distance moyenne entre les observations des clusters (ou inertie intra-cluster). Plus formellement, $\Delta(c_i, c_j) = 0 \Rightarrow c_i = c_j$ mais la réciproque est fausse.

Lorsque les données traitées sont décrites par des variables nominales et que ces dernières ont été transformées par codage disjonctif complet, les observations

deviennent des vecteurs booléens, et les clusters peuvent être représentés par des lois de probabilités. Dans ce cas, il est possible d'utiliser des divergences en guise de dissimilarités entre clusters. La plus couramment utilisée est la *divergence de Kullback-Leibler* (Cover et Thomas, 2006). Il est possible d'utiliser cette divergence sous sa forme symétrisée pour comparer deux clusters. L'inconvénient est qu'elle n'est pas bornée et qu'il suffit qu'un des deux clusters prenne la valeur 0 sur une des variables pour que sa dissimilarité avec n'importe quel autre cluster y prenant valeur soit infinie. La *divergence de Jensen-Shannon* ne pose pas ce problème et est donc plus adaptée pour comparer directement deux lois. Ces divergences sont détaillées dans le chapitre 4.

Propriétés des dissimilarités On a vu que les mesures de dissimilarités n'étaient pas nécessairement des distances. D'ailleurs, il n'existe pas de propriétés universelles pour définir une dissimilarité. Néanmoins certaines caractéristiques sont préférables pour rendre une mesure de dissimilarité exploitable (Jain et Dubes, 1988) :

- **nulle pour deux objets identiques** : qu'ils s'agissent de clusters ou d'observations, s'ils sont identiques il est préférable que leur dissimilarité soit nulle ;
- **symétrique** : de part sa définition, une dissimilarité mesure les différence entre objets et non pas les différences de l'un par rapport à l'autre. On préfère donc les mesures symétriques. Notons qu'une simple moyenne permet de symétriser des mesures asymétriques ;
- **positive** : à l'image des distances, il n'est pas intuitif d'expliquer une dissimilarité négative. On se limite donc à des définitions positives des dissimilarités.

La notion d'inertie L'inertie est un indicateur assimilable à une variance. Elle quantifie la dispersion des observations ou des clusters dans l'espace des données :

- **l'inertie inter-clusters** : l'inertie inter-clusters est une mesure qui permet de quantifier la dispersion des clusters. Il s'agit de la dissimilarité moyenne entre les clusters, pondérée par leurs populations. Dans la plupart des problèmes de clustering basés sur une mesure de similarité, cette mesure est maximisée. En effet le but d'un clustering est d'obtenir une représentation synthétique des données telle que tous les clusters soient les moins similaires les uns par rapport aux autres ;
- **l'inertie intra-cluster** : on étudie la dispersion des éléments qui composent chacun des clusters, c'est-à-dire la distance moyenne entre observations d'un même cluster. Dans ce cas là, on cherche à minimiser l'inertie de manière à obtenir des clusters groupant des valeurs les plus similaires possible ;
- **l'inertie totale** : il s'agit de la somme des inerties inter- et intra- clusters. Elle correspond à la dispersion des observations dans l'espace des données.

Elle est donc indépendante du clustering. Dans le cas où la mesure de dissimilarité utilisée respecte le théorème de Huyghens, l'inertie totale est un terme constant, correspondant à la somme des inerties inter- et intra-cluster. Cette propriété permet de minimiser l'inertie intra-cluster en maximisant l'inertie inter-clusters.

2.1.3 Choix du nombre de clusters

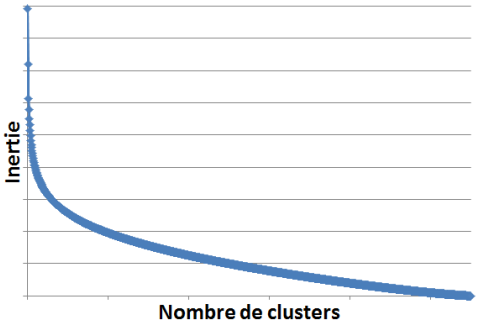
La majorité des approches de clustering sont basée sur l'optimisation d'un critère dérivé d'une mesure de dissimilarité. Peu importe la mesure choisie, le clustering qui maximise l'inertie inter-clusters et minimise l'inertie intra-cluster est toujours le clustering avec autant de clusters que d'observations (inertie intra-cluster nulle). Le choix du nombre de clusters est donc un des problèmes majeurs dans les approches de clustering.

Le nombre de clusters en tant que paramètre Dans certaines approches, le nombre de clusters est un paramètre choisi par l'utilisateur. Dans certains cas de figure, l'utilisateur connaît le nombre de clusters qu'il souhaite obtenir et donc le choix du nombre de clusters n'est pas un problème en soi. Dans le cas contraire, l'utilisateur a besoin d'outils lui permettant de faire un choix.

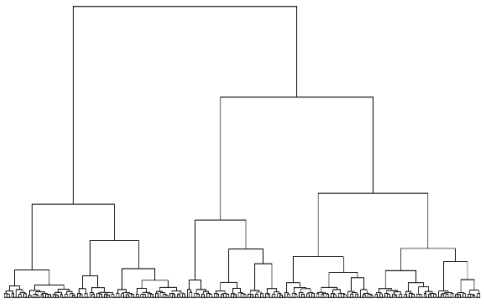
Une des solutions est d'étudier l'évolution de mesures de qualité globales du clustering – souvent liées aux inerties (Milligan et Cooper, 1985) – du meilleur clustering trouvé en fonction du nombre de clusters donné en paramètre (voir exemple de la figure 2.2). L'inertie intra-cluster diminue avec le nombre de clusters construits : le clustering ayant un unique cluster maximise la valeur de la mesure, alors que le clustering avec autant de clusters que d'observations la minimise. On cherche donc dans l'évolution le point d'inflexion permettant de trouver un bon compromis entre un clustering fin et précis et un clustering simple. Le problème de ce type d'approches est que l'évolution n'est pas forcement strictement décroissante avec un point d'inflexion marqué dans l'évolution de la mesure choisie. D'autres mesures ont été proposées afin de faciliter le choix du nombre de clusters, voir par exemple Tibshirani *et al.* (2001).

De manière similaire, l'utilisation de *dendrogrammes* permet de choisir le nombre de clusters (Hastie *et al.*, 2009) (voir exemple de la figure 2.2). Le dendrogramme est un arbre construit suite à un clustering hiérarchique. Il présente la hiérarchie des clusters depuis le clustering avec une observation par cluster jusqu'à un cluster avec toutes les observations. La hauteur des branches de l'arbre quantifie la dissimilarité entre les deux clusters fusionnés. L'utilisateur a ainsi la possibilité de choisir le niveau de grain de clustering qui lui convient.

Le nombre de clusters déterminé automatiquement La majorité des méthodes de clustering, dans lesquelles le nombre de clusters est déterminé



(a) Courbe de l'inertie intra-cluster en fonction du nombre de clusters



(b) Dendrogramme

FIGURE 2.2 – Classification hiérarchique des textes de la base *20 Newsgroups*. La courbe de l'inertie intra-cluster et le dendrogramme ont été construits pour le même clustering hiérarchique.

automatiquement, se ramènent à des problèmes de sélection de modèle (Jain, 2010). Dans ce cas, le modèle définit l'ensemble des paramètres décrivant le clustering, comme par exemple le nombre de clusters, leur composition, etc. Ces paramètres sont optimisés de manière à ce que les clusters soient les plus fidèles aux données. Par exemple, un cluster groupant des observations très similaires est une bonne représentation des observations qu'il groupe. Cette mesure de « justesse » du clustering, lorsqu'elle prend un sens statistique, est la *vraisemblance du modèle*.

Cette vraisemblance est maximale lorsque les clusters décrivent exactement les données, c'est-à-dire lorsqu'il y a autant de clusters que d'observations. Cependant, l'objectif du clustering est d'avoir un nombre suffisamment petit de clusters pour synthétiser les données. Une *pénalisation* de la vraisemblance est donc nécessaire. Cette pénalisation a pour but d'apporter un compromis entre un clustering fin et un clustering synthétique. Plusieurs approches ont été développées pour fixer cette pénalisation, parmi lesquelles on peut citer le critère d'information d'Akaike (AIC, Akaike (1974)), le critère d'information bayésienne (BIC, Fraley et Raftery (1998)), la longueur de message minimal (MML, Wallace et Boulton (1968)) ou encore la longueur de description minimale (MDL, Rissanen (1978)).

Le principal problème de la sélection de modèles est la complexité algorithmique. L'espace des modèles (ensemble de tous les clusterings possibles) est en général très grand et le critère optimisé peut admettre plusieurs optima locaux. L'utilisation d'heuristiques permettant de trouver une solution satisfaisante au problème est alors nécessaire.

2.1.4 L'évaluation du clustering

De par sa définition, un clustering est difficile à évaluer. Lorsqu'on applique un clustering sur un jeu de données, on cherche à découvrir la structure qui les caractérise. Mais sans a priori sur les données, comment évaluer la qualité du résultat obtenu ? Bien sûr, l'utilisateur peut juger satisfaisant ou non le résultat mais rien ne lui indique si le résultat est pertinent ou pas. Plusieurs solutions permettent néanmoins d'évaluer la qualité de l'algorithme.

L'évaluation sur données artificielles Cette évaluation consiste à construire un jeu de données artificiel et à appliquer un algorithme de clustering. Un bon algorithme doit permettre de retrouver les classes engendrées, à condition que les données soient effectivement structurées. Cette technique d'évaluation est illustrée dans le chapitre 2. Une bonne propriété d'un algorithme est sa *généricité*, c'est-à-dire sa capacité à trouver des clusters engendrés suivant différentes hypothèses de distributions. Pour montrer cette caractéristique, une solution est de construire des données respectant des hypothèses de distributions différentes. Un bon algorithme est alors capable de retrouver le clustering engendré quelles que soient les hypothèses de construction des données.

Un autre problème se pose : le choix du critère de qualité du clustering. Une manière classique d'évaluer la qualité des clusters retrouvés est la *pureté*. Il s'agit de traiter le clustering comme un problème supervisé. Chaque cluster engendré est associé à la classe avec laquelle il partage le plus d'observations.

Définition 4. *Soient C l'ensemble des k clusters engendrés et $\hat{C}$ l'ensemble des $\hat{k}$ clusters trouvés par l'algorithme. Pour un ensemble de m observations, la pureté est définie de la manière suivante :*

$$pureté = \frac{1}{m} \sum_{i=1}^{k} \max_{j=1..\hat{k}} (|C_i \cap \hat{C}_j|) \tag{2.4}$$

Prenons un exemple simple. Les données de la figure 2.3 sont constituées de 18 observations. Il y a trois clusters engendrés artificiellement dont les observations sont matérialisées par trois couleurs (rouge, jaune et cyan). Nous avons également trois clusters trouvés par une méthode de clustering quelconque. Le premier cluster trouvé par l'algorithme contient trois observations jaunes, le second cinq observations rouges et le troisième quatre observations cyan. On a donc une pureté de 0,66.

Le problème de la pureté est qu'un clustering avec autant de clusters que d'observations a une pureté maximale.

L'indice de Rand, noté RI, est une mesure également utilisée pour évaluer un clustering (Rand, 1971). On considère deux partitions des données en clusters : C l'ensemble des k clusters engendrés et $\hat{C}$ l'ensemble des $\hat{k}$ clusters trouvés par l'algorithme. On cherche à grouper deux observations dans un même cluster si elle proviennent d'un même cluster artificiel. On mesure donc le taux de bonne

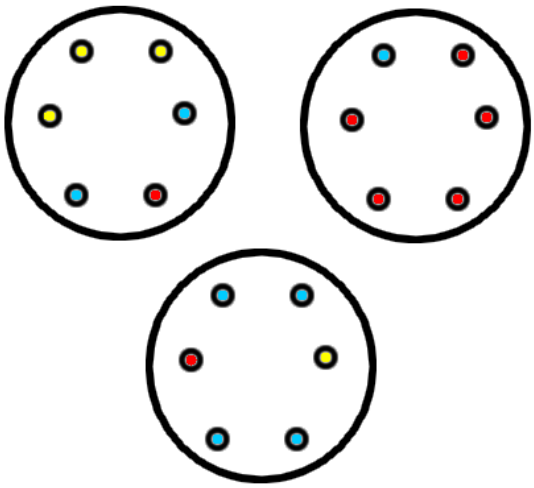


FIGURE 2.3 – Exemple de clustering. Les cercles noirs représentent des clusters obtenus via une approche de clustering. Les couleurs modélisent l'appartenance à des classes engendrées artificiellement.

assignation de ces paires d'observations. Pour chaque couple d'observations, on peut compter quatre types d'assignations possibles :

- les vrais positifs (VP) : les deux observations sont dans le même cluster dans C et $\hat{C}$,
- les vrais négatifs (VN) : les deux observations sont dans des clusters différents dans C et $\hat{C}$,
- les faux positifs (FP) : l'algorithme a classé les observations dans le même cluster dans $\hat{C}$ alors qu'elles ne le sont pas dans C,
- les faux négatifs (FN) : l'algorithme a classé les observations dans des clusters différents de $\hat{C}$ alors qu'elles sont dans le même dans C.

Définition 5. *L'indice de Rand est défini comme le pourcentage de bonnes assignations.*

$$RI = \frac{|VP| + |VN|}{|VP| + |VN| + |FP| + |FN|} \tag{2.5}$$

Le nombre total d'assignations ($|VP| + |VN| + |FP| + |FN|$) *vaut* $\frac{1}{2}m(m-1)$ *où* m *est le nombre d'observations.*

L'indice de Rand est compris entre 0 et 1. Deux clustering exactement identiques donnent une valeur de 1 à la mesure. Le problème de l'indice de Rand est que pour deux partitions aléatoires, on peut avoir des valeurs différentes et non nulles. Une version ajustée ARI de cette mesure est donc plus souvent utilisée : l'*indice de Rand Ajusté* (Hubert et Arabie, 1985). On compare la valeur de l'indice de Rand avec sa valeur attendue en cas d'indépendance de la classe engendrée et de la classe retrouvée.

Définition 6. *L'indice de Rand peut être ajusté de la manière suivante :*

$$ARI = \frac{RI - ERI}{max(RI) - ERI} \tag{2.6}$$

où ERI est la valeur de l'indice de Rand avec sa valeur attendue en cas d'indépendance de la classe engendrée et de la classe retrouvée.

Reprenons l'exemple de la figure 2.3, nous avons 153 assignations possibles, 20 sont des vrais positifs et 80 sont des vrais négatifs. On a donc un indice de Rand de 0,65 et une indice de Rand ajusté de 0,18.

L'information mutuelle est une autre mesure d'évaluation basée sur la théorie de l'information qui permet de mesurer la corrélation entre la répartition des observations dans les clusters engendrés artificiellement et la répartition des observations dans les clusters trouvés par l'algorithme. Sa version normalisée (Strehl et Ghosh, 2003) est plus souvent utilisée afin de rendre la mesure plus interprétable : s'il y a identité entre les clusters engendrés et les clusters trouvés, cette mesure vaut 1.

Définition 7. *Soient C l'ensemble des k clusters engendrés et $\hat{C}$ l'ensemble des $\hat{k}$ clusters trouvés par l'algorithme. L'information mutuelle est définie de la manière suivante :*

$$MI(C, \hat{C}) = \sum_{i=1}^{k} \sum_{j=1}^{\hat{k}} p(C_i, \hat{C}_j) \log \frac{p(C_i, \hat{C}_j)}{p(C_i)p(\hat{C}_j)} \tag{2.7}$$

L'information mutuelle admet un maximum égal à la moyenne des entropies H des partitions engendrées et trouvée. Ce maximum est donc utilisé pour normaliser l'information mutuelle :

$$NMI(C, \hat{C}) = \frac{2MI(C, \hat{C})}{H(C) + H(\hat{C})} \tag{2.8}$$

Reprenons l'exemple de la figure 2.3, l'information mutuelle entre la partition engendrée artificiellement et la partition inférée vaut 0,29, sa version normalisée vaut 0,27.

Pour résumer, ces mesures permettent de vérifier que l'approche de clustering regroupe des observations qui ont été engendrées de manière à provenir d'un même groupe.

2.1.5 Tendance au clustering

La tendance au clustering consiste à déterminer a priori si les données peuvent être partitionnées en clusters. Les approches de clustering basées sur l'optimisation d'un critère non régularisé, construit sur une mesure de similarité, produisent des clusters, même si les données sont distribuées de

manière aléatoire. On a vu dans la section 2.1.4 qu'il est possible d'évaluer un algorithme de clustering sur des bases de données artificielles, mais l'évaluation d'un clustering sur une base de données réelle est subjective. On cherche donc à éviter une utilisation inappropriée du clustering sur des données ne présentant aucune structure sous-jacente. Ainsi, tester la tendance au clustering permet d'éviter l'application inutile d'un algorithme de clustering et d'une étape d'évaluation des résultats (Jain et Dubes, 1988). De plus, en procédant ainsi on évite de produire des résultats non pertinents, ce qui augmente la fiabilité dans la démarche d'analyse exploratoire.

Des données ayant une tendance au clustering sont distribuées de manière non uniforme sur leur espace de définition. La tâche principale du test de tendance au clustering se ramène donc à déterminer si les données sont uniformément distribuées, en amont du clustering. Une technique consiste à déterminer l'indice d'Hopkins (Diggle, 1983). Pour le déterminer : (i) on calcule la dissimilarité moyenne entre les observations et leur plus proche voisin, (ii) on engendre uniformément autant de points qu'il y en a dans les données et dans l'espace dans lequel sont définis les observations, (iii) on calcule la dissimilarité moyenne entre les observations et leur plus proche voisin aléatoirement engendré. L'indice d'Hopkins est ensuite calculé.

Définition 8. *Soit $\mathcal{D} = \{u^{(i)}\}_{i=1..m}$ l'ensemble des m observations présentes dans les données. $\mathcal{D}^A = \{v^{(i)}\}_{i=1..m}$ est l'ensemble des données générées aléatoirement, définies sur le même espace que $\mathcal{D}$. La dissimilarité entre une observation et son plus proche voisin dans $\mathcal{D}$ est notée $w_i = \min_{\substack{j=1..m \\ j\neq i}} \big(D(u^{(i)}, u^{(j)})\big)$. La dissimilarité entre une observation et son plus proche voisin dans $\mathcal{D}^A$ est notée $q_i = \min_{\substack{j=1..m \\ j\neq i}} \big(D(u^{(i)}, v^{(j)})\big)$. L'indice d'Hopkins est défini par :*

$$H = \frac{\sum_{i=1}^{m} q_i}{\sum_{i=1}^{m} w_i + \sum_{i=1}^{m} q_i} \tag{2.9}$$

Si les données ne sont pas disposées au clustering, alors l'index est proche de 0,5. S'il y a présence de groupes naturels d'observations et donc une tendance au clustering, alors la dissimilarité moyenne entre les observations et leur plus proche voisin artificiel est plus importante que la dissimilarité moyenne entre les observations et leur plus proche voisin dans les données. Dans ce cas, l'indice d'Hopkins est compris entre 0,5 et 1, ce dernier étant d'autant plus proche de 1 que les données se prêtent au clustering.

Une seconde approche (Jain et Dubes, 1988) consiste à construire un graphe dont les nœuds modélisent les observations des données et des observations artificielles générées de la même manière que pour le calcul de l'indice d'Hopkins. Ces nœuds sont reliés par des arêtes pondérées par la dissimilarité entre les observations. L'arbre couvrant de poids minimal est calculé sur ce graphe. Il s'agit du sous-graphe qui connecte tous les nœuds du graphe de manière à ce que la somme des poids des arêtes soit minimale. Une fois l'arbre construit, on

compte le nombre d'arêtes connectant un nœud issu des données réelles et un nœud issu des données artificielles. Plus ce nombre est faible, plus les données ont une tendance au clustering.

De nombreuses techniques basées sur des tests statistiques sont détaillées par Jain et Dubes (1988). Des approches basées sur des interprétations visuelles ont également été proposées par Bezdek et Hathaway (2002).

2.1.6 L'exploitabilité du clustering

Lorsque le clustering est réalisé sur un jeu de données réelles, la pertinence du résultat est subjective et liée à un jugement de l'utilisateur. L'exploitabilité des résultats peut malgré tout être étudiée. Une bonne approche de clustering n'est pas simplement une approche fiable, c'est surtout une approche qui permet de tirer une information pertinente de données présentant une structure sous-jacente. Lorsque le volume des données est important, le problème se complexifie : plus il y a de données, plus il y a potentiellement de clusters. C'est pourquoi il est important de permettre à l'utilisateur d'extraire une information utile et synthétique du clustering.

Les approches de clustering hiérarchique permettent à l'utilisateur de régler le niveau de précision souhaité dans le clustering. Cette hiérarchie peut être construite pour les méthodes de clustering basée sur une sélection de modèle. Pour cela, on calcule la matrice de dissimilarité inter-clusters depuis le nombre de clusters trouvé par l'algorithme jusqu'à la racine de la hiérarchie, c'est-à-dire le modèle avec un seul cluster. Cette façon de procéder est coûteuse mais permet à l'utilisateur de sélectionner un niveau de précision qui le satisfait, du plus précis au plus synthétique. De plus, l'avantage sur une classification hiérarchique classique est qu'on ne risque pas de choisir un modèle avec des clusters non significatifs.

Lorsque l'approche de clustering produit un trop grand nombre de clusters mais qu'on souhaite malgré tout conserver le niveau de précision maximal, on cherche à se focaliser sur les clusters les plus intéressants. Le problème est de mesurer la notion d'*intérêt* des clusters. Les premières définitions de l'intérêt des clusters ont été principalement subjectives, nécessitant des connaissances a priori des caractéristiques des clusters intéressants que l'on souhaite observer (Silberschatz et Tuzhilin, 1995). Plus récemment, des mesures de l'intérêt ont été construites selon des principes de théorie de l'information (De Bie, 2011). Mais ces dernières sont également subjectives car elles dépendent d'hyperparamètres fournis par l'utilisateur. Geng et Hamilton (2006) ont fait une liste des caractéristiques que doit posséder un cluster intéressant. Un cluster intéressant doit être simple à analyser par l'utilisateur. Il doit être surprenant dans le sens où le cluster trouvé va à l'encontre de l'a priori que peut avoir l'utilisateur sur les données. Il doit grouper des observations qui soient suffisamment nombreuses pour que le cluster soit significatif et éloignées des autres observations pour que le cluster soit atypique.

2.2 Le co-clustering, définitions et état de l'art

Le co-clustering est une technique qui a pour but de réaliser une partition simultanée des lignes et des colonnes d'une matrice de données (Hartigan, 1972; Mirkin, 1996). On distingue trois types de co-clustering (Van Mechelen *et al.*, 2004). Dans le cas classique, les dimensions de la matrice sont les observations et les variables. Dans ce cas, on réalise simultanément une partition des individus et des variables descriptives des données. Le co-clustering peut également consister à réaliser une segmentation conjointe des valeurs de deux variables descriptives des observations. Cette catégorie de co-clustering est différente des approches initiales. Ici on réalise une partition d'une table de contingence. Enfin, le co-clustering peut aussi être appliqué sur une matrice de similarité entre deux ensembles d'observations, identiques ou différents. L'approche MODL, sur laquelle est basée cette thèse, permet de réaliser un co-clustering des valeurs de d variables descriptives des données. On choisit donc une notation adaptée à ce type d'approches de co-clustering.

Les données étudiées dans les problèmes de co-clustering sont de même nature que les données traitées par les approches de clustering : elles sont composées de m *observations* sans étiquette (ou *classe*), décrites par plusieurs *variables*, notées $\{X_1, X_2, ..., X_d\}$. Ces variables peuvent être continues ou nominales, prenant alors un nombre fini n_j de *valeurs* différentes. On cherche à partitionner les valeurs prises par les variables descriptives afin d'obtenir de nouvelles variables $\{X_1^M, X_2^M, ..., X_d^M\}$, que nous appelons *variables-partitions*. Les modalités de ces variables sont les clusters obtenus par les partitions des valeurs des variables $\{X_1, X_2, ..., X_d\}$. Chacune de ces variables a $\{k_1, k_2, ..., k_d\}$ modalités qui sont des groupes de valeurs si la variable est nominale et des intervalles si la variable est continue.

Dans cette section, plusieurs approches sont succinctement présentées (voir Madeira et Olivieira (2004), Charrad et Ben Ahmed (2011) et Govaert et Nadif (2013) pour des états de l'art complets sur le co-clustering), dont l'approche MODL (Boullé, 2007) qui servira de base pour le reste de la thèse.

2.2.1 Les premières approches

Les premières approches de co-clustering sont des partitions conjointes des observations et des variables. La première approche est souvent attribuée à Hartigan (1972). Ce dernier n'emploie cependant pas le terme de co-cluster pour désigner une sous-matrice dans une matrice de contingence, mais le terme cluster. Il utilise l'exemple des résultats des votes des assemblées générales des Nations Unies tenues en 1969 et 1970. Les données sont traitées sous leur forme tabulaire : les pays votant en ligne, les résolutions soumises au vote en colonne et dans les cases de la matrice le résultat du vote (oui, non, abstention, absent du vote). L'intérêt de faire simultanément des clusters de pays votants et de résolutions est alors souligné par Hartigan (1972), ce qu'il met en application en faisant émerger des clusters de pays du bloc de l'Est, de l'Ouest et du Tiers

Monde en parallèle de clusters de résolutions.

Hartigan (1972) propose de construire une matrice A^* modélisant l'interaction moyenne des clusters de lignes et de colonnes de la matrice de contingence étudiée A. La mesure de qualité de la matrice A^* par rapport à la matrice A est la somme des carrés résiduels $\sum_{i,j}(A_{ij} - A^*_{ij})^2$ qui est minimisée. L'algorithme proposé est hiérarchique descendant. À l'initialisation, on a donc un seul bloc et tous les éléments la matrice A^* sont égaux à la moyenne des éléments de la matrice A et la somme des carrés résiduels équivaut à la variance des éléments de la matrice A. Au contraire, lorsqu'il y a autant de clusters que de lignes et de colonnes, A^* est identique à A et la somme des carrés résiduels vaut 0. À chaque étape de l'algorithme, la partition d'une sous-matrice de A en deux sous-matrices maximisant la perte du critère est opérée. On cherche donc à réaliser la partition réduisant au maximum la variance au sein des sous-matrices. À chaque étape, la perte de critère est également calculée pour des partitions aléatoires. Lorsque la meilleure partition n'est pas meilleure que la partition aléatoire, l'algorithme est arrêté.

On obtient donc une partition des lignes et des colonnes ainsi qu'une hiérarchie des clusters, à l'image des méthodes de clustering hiérarchique classiques. Des méthodes similaires cherchant le découpage optimal de matrices de données en sous-matrices ont également été développées par Eckes et Orlik (1993) et Mirkin *et al.* (1995).

D'autres approches assimilables à du co-clustering consistent à ordonner les lignes et les colonnes des matrices de données, de manière à grouper les lignes et les colonnes similaires. Ces clusters peuvent être obtenus en utilisant une mesure de similarité comme la mesure de χ^2 (Govaert, 1977). Ils peuvent également être obtenus en cherchant des blocs denses et/ou homogènes dans la matrice. On parle dans ce cas de *Blockmodeling* (Breiger *et al.*, 1975; Arabie *et al.*, 1988). Issus principalement de concepts sociométriques, cette notion de blockmodeling est étudiée plus en détails dans la section 3.2 traitant de l'application du co-clustering au partitionnement de graphe et à l'analyse de réseaux sociaux.

Comme dans certains algorithmes de clustering, l'optimisation de l'inertie a également été utilisée dans certaines approches de co-clustering. Govaert (1995) propose un algorithme agglomératif permettant de maîtriser la dégradation de l'inertie des clusters à chaque fusion de lignes ou de colonnes. Le critère optimisé par cette approche peut être interprété comme une généralisation du critère des k-means pour les données tabulaires (Van Mechelen *et al.*, 2004). Notons que l'approche d'Hartigan (1972) minimisant la variance dans les blocs de la matrice de données peut également être interprétée comme une méthode d'optimisation de l'inertie associée à la sommes des carrés résiduels en tant que mesure de dissimilarité.

Mirkin (1996) dans un état de l'art des techniques de clustering introduit

pour la première fois le terme de *biclustering* en tant que clustering simultané des lignes et des colonnes dans les données matricielles. Cheng et Church (2000) ont notamment illustré l'intérêt du co-clustering dans le domaine de la biostatistique et plus particulièrement dans les problèmes d'expression de gènes. De manière similaire à l'approche d'Hartigan (1972), le co-clustering est traité comme un problème d'optimisation : chaque bicluster est considéré comme une sous-matrice et associé à un *résidu quadratique moyen* qui est minimisé afin d'obtenir de grands biclusters homogènes.

2.2.2 Les approches stochastiques

Dans les approches dites stochastiques l'hypothèse est faite que les données sont issues d'un mélange de lois sous-jacent aux données.

Dans les mélanges finis, on considère que les données sont engendrées suivant un mélange de k lois modélisant les clusters. On peut donc définir un modèle décrit par l'ensemble des k paramètres de distribution $\theta = \{\theta_1 ... \theta_k\}$ des observations dans les clusters et des coefficients de mélange des k clusters $\pi = \{\pi_1 ... \pi_k\}$. On cherche donc à maximiser la vraisemblance $p(\mathcal{D}|\pi, \theta)$ d'observer les données connaissant les paramètres du modèle. On introduit une variable latente z indiquant l'appartenance des observations à chacun des clusters. Les m observations $u^{(i)}$ sont supposées avoir été engendrées indépendamment.

Définition 9. *La probabilité d'observer les données connaissant les paramètres π et θ est définie de la manière suivante :*

$$\begin{aligned} p(\mathcal{D}|\pi, \theta) &= \prod_{i=1}^{m} \sum_{j=1}^{k} \pi_j p(u^{(i)}|\theta_j) \\ &= \sum_{z=1}^{k} p(z|\pi) p(\mathcal{D}|\theta, z) \end{aligned} \tag{2.10}$$

Dans le cas où la loi est normale, le paramètre θ est l'ensemble des moyennes et des variances de chacune des lois associées aux clusters. Ces paramètres peuvent être inférés en utilisant des algorithmes de type Espérance-Maximisation (EM).

Dans le cas du co-clustering, les données observées dans une table de contingence. On cherche une partition conjointe des lignes et des colonnes.

Définition 10. *Govaert et Nadif (2003) proposent un modèle de mélange adapté au biclustering où deux partitions sont calculées simultanément :*

$$p(\mathcal{D}|\pi, \theta) = \sum_{z^{(1)}=1}^{k_1} \sum_{z^{(2)}=1}^{k_2} p(z^{(1)}|\pi^{(1)}) p(z^{(2)}|\pi^{(2)}) p(\mathcal{D}|\theta, z^{(1)}, z^{(2)}) \tag{2.11}$$

où $z^{(1)}$ et $z^{(2)}$ sont respectivement les variables latentes associées aux lignes et aux colonnes de la matrice de données, et $\pi^{(1)}$ et $\pi^{(2)}$ les coefficients de mélanges des clusters associés aux partitions.

Les probabilités a priori sont déterminées indépendamment sur chacune des variables alors que la vraisemblance est définie comme la probabilité d'observer les données connaissant les paramètres de co-clustering. Plusieurs approches permettant de déterminer ces paramètres ont été proposées en utilisant des approches de type EM (Govaert et Nadif, 2003; Nadif et Govaert, 2010) ou encore des méthodes variationnelles (Shan et Banerjee, 2008).

Un modèle génératif populaire est l'*allocation latente de Dirichlet (LDA)* (Blei *et al.*, 2003). Ce modèle a été initialement introduit dans le cadre de la classification de documents. Il s'agit de faire un clustering des mots afin d'obtenir une description simplifiée de chaque document. Pour déterminer le nombre de clusters et la proportion d'éléments qu'ils contiennent, une loi a priori de Dirichlet de paramètre α pénalise la vraisemblance du modèle.

Définition 11. *La fonction de densité de probabilité associée au modèle LDA est définie de la manière suivante :*

$$p(\mathcal{D}|\alpha,\theta) = \int_{\pi} Dir(\pi|\alpha) \left(\prod_{i=1}^{m} \sum_{z_i=1}^{k} p(z_i|\pi) p(u^{(i)}|\theta, z_i) \right) d\pi \tag{2.12}$$

Cette probabilité peut être estimée par des méthodes variationnelles (Jordan *et al.*, 1999) ou encore par un échantillonnage de Gibbs (Gelfand et Smith, 1990). Il ne s'agit pas directement d'une approche de co-clustering mais peut être adaptée afin de réaliser un clustering conjoint des textes et des mots (Shan et Banerjee, 2008).

Dans le domaine de la sociométrie, Holland *et al.* (1983) ont introduit la notion de blockmodeling stochastique. Différentes approches stochastiques ont par la suite été proposées, elles sont détaillées dans la section 3.2 dédiée à l'utilisation du co-clustering pour l'analyse de graphes.

L'avantage des approches de co-clustering stochastique est de pouvoir traiter des données nominales et continues (Nadif et Govaert, 2010) et donc d'envisager un large ensemble d'applications.

2.2.3 Les approches spectrales

De manière générale, le clustering spectral consiste à prendre une matrice de similarité entre observations et à la réduire de manière à construire des clusters. Notons A la matrice de similarité de taille $m \times m$, m étant le nombre d'observations. La matrice D est une matrice diagonale telle que $D_{ii} = \sum_{j=1}^{m} A_{ij}$. On peut alors définir la matrice Laplacienne normalisée L telle que :

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

Les k plus grandes valeurs propres de la matrice L sont sélectionnées et une nouvelle matrice X de taille $m \times k$ est construite. Les vecteurs propres normalisés par leur norme euclidienne forment les colonnes de la matrice X. Enfin en appliquant un algorithme de clustering (hiérarchique ou k-means par exemple)

sur chacune des lignes de la matrices X modélisant les m observations de départ dans un espace $\mathbb{R}^k$, on obtient un ensemble de clusters (Ng *et al.*, 2002).

Dhillon (2001) propose une approche spectrale pour le biclustering adaptée à la classification de textes et de mots. Une matrice A est construite avec en ligne les mots, en colonne les documents et dans les cases la fréquence d'apparition d'un mot dans un document. Une version normalisée A_n de la matrice de co-occurrences est définie par :

$$A_n = D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}}$$

D_1 (resp. D_2) est la matrice diagonale dans laquelle les éléments diagonaux sont la fréquence des mots (resp. la taille des textes). Au lieu de calculer les vecteurs propres, les vecteurs singuliers de gauche et de droite, correspondant à la seconde plus importante valeur singulière, sont calculés. La partition des mots et des documents est obtenue de la même manière que pour le clustering spectral calculé avec les vecteurs propres.

Notons que l'exemple du co-clustering de textes et de mots peut être vu comme un co-clustering observations/variables ou modalités/modalités. En effet, on peut considérer les textes comme les textes comme les observations et les mots comme des variables. On peut également voir le problème de classification de documents comme une partition de matrice de contingence où les dimensions sont une variable mot et une variable texte. Dans ce cas les éléments de la matrices sont les fréquences ou les probabilités jointes des textes et des mots.

Kluger *et al.* (2003) proposent une extension de l'approche de Dhillon (2001) pour le biclustering adaptée aux problèmes d'expressions de gènes.

2.2.4 Les approches basées sur la théorie de l'information

L'information mutuelle est une mesure fréquemment utilisée dans l'évaluation du clustering de manière générale (voir section 2.1.4). Elle permet de mettre en évidence les corrélations entre deux variables. Quand elle est utilisée dans le cadre d'une mesure d'évaluation sur des données artificielles dont les clusters sont connus, elle étudie les corrélations entre les clusters engendrés et les clusters retrouvés par une approche.

Dans le cadre du co-clustering en deux dimensions, Dhillon *et al.* (2003) proposent un critère minimisant la perte d'information mutuelle lorsque les valeurs des deux variables sont partitionnées. En effet, le partitionnement des valeurs des variables engendre nécessairement une perte d'information. Ainsi l'information mutuelle entre les deux variables descriptives des données est nécessairement supérieure à l'information mutuelle entre les deux variables issues du partitionnement simultané des deux dimensions de la matrice de données.

Définition 12. *Soit un ensemble de données $\mathcal{D}$ décrites par deux variables X_1 et X_2. On cherche un modèle $\mathcal{M}$ de biclustering, définit par les partitions*

conjointes des valeurs des variables X_1 et X_2. Les variables-partitions X_1^M et X_2^M prennent valeurs dans l'ensemble des clusters obtenus respectivement sur les variables X_1 et X_2. Le biclustering est optimal si la perte d'information mutuelle est minimisée.

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmin}}(MI(X_1, X_2) - MI(X_1^M, X_2^M)) \tag{2.13}$$

En démontrant que la perte d'information mutuelle peut se décomposer sous forme d'une somme pondérée de divergences de Kullback-Leibler entre la loi de probabilité des lignes conditionnellement aux clusters des colonnes et la loi des colonnes conditionnellement aux clusters des lignes, Dhillon *et al.* (2003) définissent un algorithme qui décroit de manière monotone jusqu'à un optimum local. Dans les problèmes de clustering de documents et de mots, cet algorithme sera particulièrement efficace du fait de sa complexité linéaire en fonction du nombre de co-occurences de documents et de mots. Une généralisation des méthodes basées sur la perte d'information d'un modèle en biclusters est proposée par Banerjee *et al.* (2004).

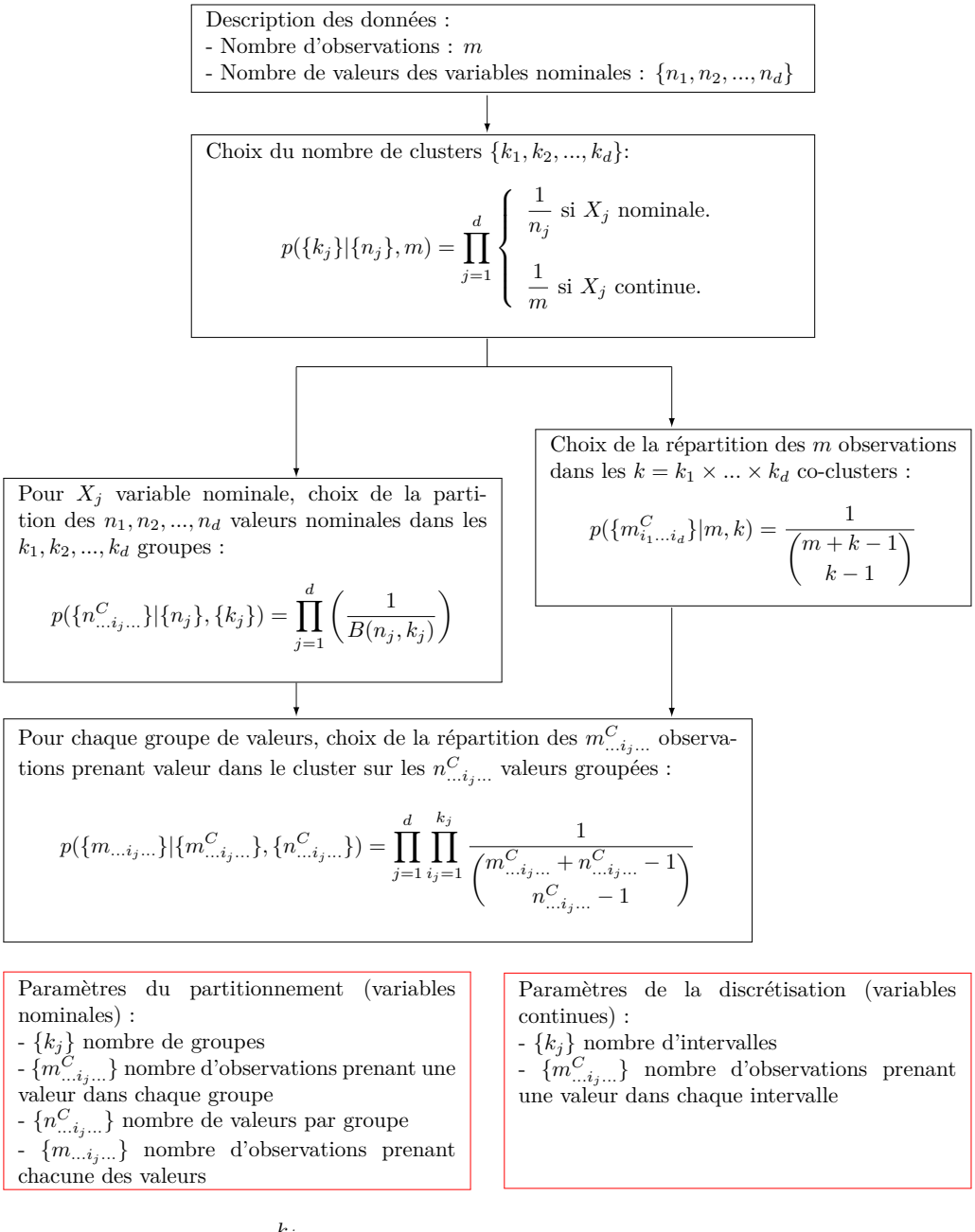
2.2.5 MODL

L'approche MODL (Boullé, 2007) est basée sur la sélection de modèle. Le meilleur modèle est sélectionné de manière à maximiser une vraisemblance pénalisée par une loi de probabilité a priori sur l'ensemble des paramètres. Ces paramètres sont organisés de manière hiérarchique afin de construire une loi a priori elle-même hiérarchique. À chaque niveau de la hiérarchie, la loi est uniforme de manière à être la moins informative possible. Contrairement à beaucoup d'approches génératives classiques, MODL ne fait pas d'hypothèse d'indépendance conditionnelle des paramètres. On cherche donc à les générer en une seule fois. En outre, la loi a priori est construite pour une valeur fixée du nombre d'observations (et de modalités). Le modèle $\mathcal{M}$ est défini par un ensemble de paramètres caractéristiques d'un co-clustering en d dimensions avec des variables continues ou nominales :

1. le nombre de clusters par dimension,
2. la partition de l'ensemble des valeurs (ou modalités) de chacune de variables nominales (pour un nombre de clusters connus),
3. l'effectif de chacun des co-clusters (en nombre d'observations),
4. le nombre observations prenant chacune des valeurs (ou modalités) dans les clusters (dans le cas de variables nominales, en respectant la contrainte induite par les effectifs des co-clusters).

Afin de rechercher le meilleur modèle, on applique une approche MAP visant à maximiser la probabilité. Nous définissons dans un premier temps

la loi a priori. La loi a priori à chaque niveau de la hiérarchie est définie conditionnellement aux valeurs des paramètres des niveaux précédents. Les choix des paramètres sont indépendants au sein d'un niveau (par exemple les nombres de clusters sont choisis indépendamment sur chaque variable). Enfin, les paramètres sont choisis de manière uniforme à chaque niveau de la hiérarchie. comme l'illustre le schéma suivant :



Le terme $B(n_j, k_j) = \sum_{\kappa=1}^{k_j} \left\{ {n_j \atop \kappa} \right\}$ est une somme de nombres de Stirling de second ordre, c'est-à-dire le nombre de partitions en κ sous-ensembles d'un ensemble de n_j éléments.

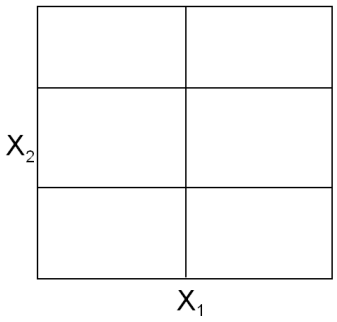
Le nombre de paramètres nécessaires pour définir les variables continues est inférieur à celui des variables nominales. En effet, il est inutile de spécifier le nombre de valeurs différentes par intervalles puisque les données numériques sont normalisées par statistique d'ordre. De ce fait, on aura une seule possibilité de distribuer les valeurs dans les intervalles puisqu'elles sont ordonnées. On aura également une seule possibilité de distribuer les observations sur les valeurs des intervalles puisque, pour chaque observation, est attribué un unique rang sur chacune des variables continues.

Définition 13. *La probabilité a priori d'un modèle $\mathcal{M}$ de co-clustering en d dimensions est définie par le produit des probabilités de ses paramètres :*

$$\begin{aligned} p(\mathcal{M}) = {} & p(\{k_j\}|\{n_j\}, m) \times p(\{m^C_{i_1 \dots i_d}\}|m, k) \\ & \times p(\{n^C_{\dots i_j \dots}\}|\{n_j\}, \{k_j\}) \times p(\{m_{\dots i_j \dots}\}|\{m^C_{\dots i_j \dots}\}, \{n^C_{\dots i_j \dots}\}) \end{aligned} \tag{2.14}$$

Prenons un exemple simple pour illustrer la construction de la loi a priori. On considère des données, composées de $m = 100$ observations décrites par deux variables X_1 (nominale) et X_2 (continue). La variable X_1 prend des valeurs dans un ensemble de $n_1 = 10$ valeurs (ou modalités). La variable X_2 prend des valeurs sur un domaine numérique quelconque. Nous détaillons la valeur des termes de la loi de probabilité a priori pour chaque niveau de la hiérarchie pour un modèle en particulier :

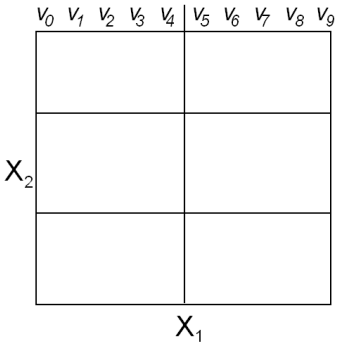
- le choix du nombre de groupes de valeurs k_1 pour la variable X_1 et du nombre d'intervalles k_2 pour la variable X_2 sont faits indépendamment. Le nombre de groupes de valeurs est compris entre 1 et 10 (le nombre de modalités pouvant être prises par X_1), tandis que le nombre d'intervalles est compris entre 1 et 100 (le nombre total d'observations).



$$p(k_1|n_1 = 10) = \frac{1}{10} \text{ et } p(k_2|m = 100) = \frac{1}{100}$$

On choisit $k_1 = 2$ et $k_2 = 3$;

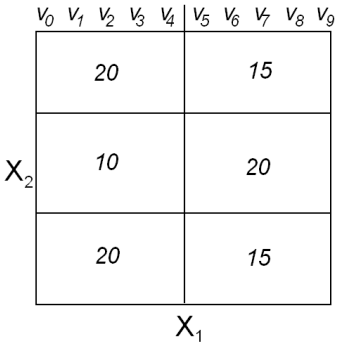
- le choix de la partition de l'ensemble des modalités prises par la variable X_1 est équiprobable parmi toutes les partitions possibles de 10 éléments dans 2 sous-ensembles, dont l'un est potentiellement vide.



$$p(\{n_{i_1.}^C\}|n_1 = 10, k_1 = 2) = \frac{1}{B(10,2)} = \frac{1}{512}$$

Nous choisissons la partition suivante : $\{v_0, v_1, v_2, v_3, v_4\}$ et $\{v_5, v_6, v_7, v_8, v_9\}$;

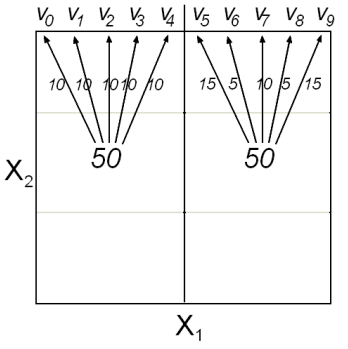
- le nombre de clusters et d'intervalles nous permet de connaître le nombre de co-clusters : $k = k_1 \times k_2 = 6$. On considère équiprobables toutes les répartitions possibles des $m = 100$ observations dans 6 co-clusters. On choisis la répartition ci-contre.



$$p(\{m_{i_1 i_2}^C\}|m = 100, k = 6) = \frac{1}{\binom{100+6-1}{6-1}} = \frac{1}{9{,}656.10^{10}}$$

Notons que les effectifs des co-clusters nous permettent de déduire le nombre d'observations prenant valeurs dans les intervalles et dans les groupes de valeurs par des sommes sur les colonnes et sur les lignes. Dans le cas de la variable X_2, les valeurs numériques étant ordonnées et pré-traitées par statistique d'ordre, on en déduit les intervalles ;

- pour chacun des deux groupes de valeurs de la variable X_1, nous répartissons les effectifs (en termes d'observations) sur les cinq valeurs groupées.



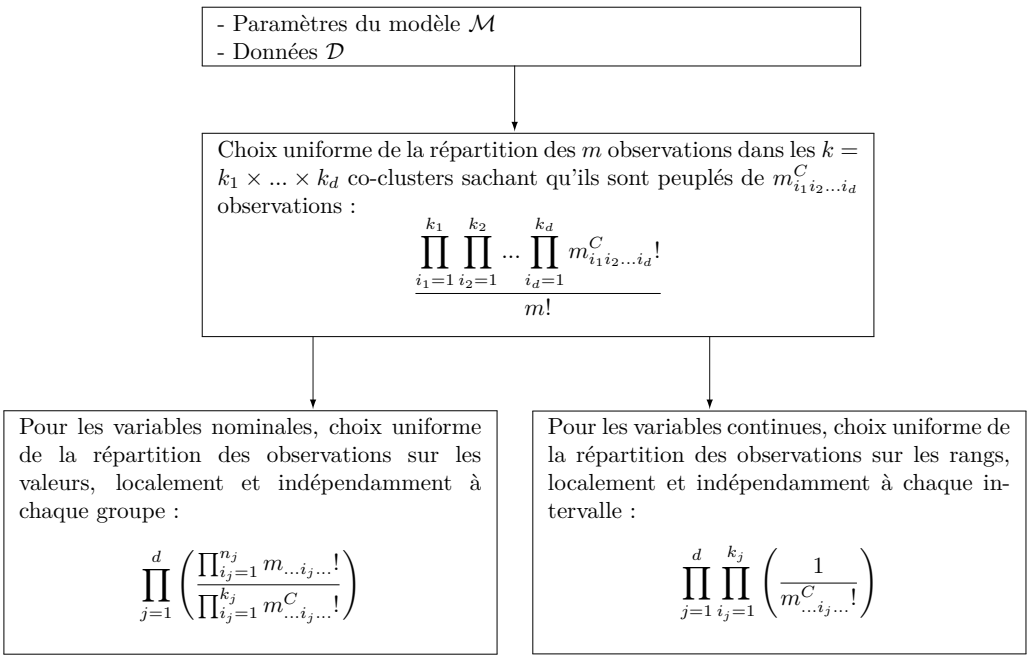
$$p(m_{0.}, m_{1.}, m_{2.}, m_{3.}, m_{4.}|m_{1.}^C = 50, n_{1.}^C = 5) = \frac{1}{\binom{50+5-1}{5-1}} = \frac{1}{3{,}163.10^5}$$

$$p(m_{5.}, m_{6.}, m_{7.}, m_{8.}, m_{9.} | m^C_{2.} = 50, n^C_{2.} = 5) = \frac{1}{\binom{50+5-1}{5-1}} = \frac{1}{3{,}163.10^5}$$

Dans le cas continu, on a fait l'hypothèse qu'il y a autant de valeurs que d'observations. Ces termes ne sont donc pas à spécifier dans la loi de probabilité a priori.

Dans cet exemple simple, nous avons un modèle de co-clustering dont le coût de codage, c'est-à-dire le logarithme négatif du produit des termes de la loi a priori, vaut 56,86.

Une fois les paramètres du modèle spécifiés, la vraisemblance est définie comme la probabilité d'observer les données connaissant les paramètres du modèle. Les observations sont distribuées en un seul tirage selon une hiérarchie et de manière uniforme à chaque niveau de cette hiérarchie. Les termes de vraisemblance de chacun des paramètres du modèle sont détaillés dans le schéma suivant :



Définition 14. *La vraisemblance $\mathcal{L}$ d'un modèle $\mathcal{M}$ de co-clustering en d dimensions est définie par le produit des termes de vraisemblance de ses paramètres :*

$$\begin{aligned} \mathcal{L}(\mathcal{M}) = & \frac{\prod_{i_1=1}^{k_1} \prod_{i_2=1}^{k_2} ... \prod_{i_d=1}^{k_d} m^C_{i_1 i_2 ... i_d}!}{m!} \\ & \times \prod_{j=1}^{d} \left(\frac{\prod_{i_j=1}^{n_j} m_{...i_j...}!}{\prod_{i_j=1}^{k_j} m^C_{...i_j...}!} \right) \times \prod_{j=1}^{d} \prod_{i_j=1}^{k_j} \left(\frac{1}{m^C_{...i_j...}!} \right) \end{aligned} \tag{2.15}$$

Reprenons l'exemple introduits pour illustrer la construction de la loi a priori. Les observations sont étiquetées : $\mathcal{D} = \{u_1, ..., u_{100}\}$. On note les clusters $c_{i,j}$ avec i et j respectivement les i^e et j^e clusters associés aux variables-partitions X_1^M et X_2^M.

- toutes les données satisfaisant la répartition des observations dans les co-clusters, spécifiée dans la loi a priori, peuvent être engendrées avec la même probabilité.

$$\frac{20!10!20!15!20!15!}{100!} = \frac{1}{1{,}04.10^{72}}$$

On choisit la configuration ci-contre parmi les $1{,}04.10^{72}$ configurations possibles ;

obs.	X_1	X_2	*co-cluster*
u_1			$c_{0,0}$
...			...
u_{20}			$c_{0,0}$
u_{21}			$c_{0,1}$
...			...
u_{30}			$c_{0,1}$
...			...
u_{100}			$c_{2,3}$

- pour la variable X_1 (nominale), toutes les données satisfaisant la répartition des observations sur les valeurs, spécifiée dans la loi a priori localement et indépendamment pour chaque groupe, peuvent être engendrées avec la même probabilité.

$$\frac{10!10!10!10!10!}{50!} = \frac{1}{4{,}83.10^{31}}$$

$$\frac{15!5!10!5!15!}{50!} = \frac{1}{3{,}40.10^{29}}$$

obs.	X_1	X_2	*co-cluster*
u_1	v_0		$c_{0,.}$
...	...		...
u_{50}	v_4		$c_{0,.}$
u_{51}	v_5		$c_{1,.}$
...	...		...
u_{100}	v_9		$c_{1,.}$

- pour la variable X_2 (continue), toutes les données satisfaisant la répartition des observations dans les intervalles spécifiée dans la loi a priori, peuvent être engendrées avec la même probabilité, sachant que la variable continue est normalisée par statistique de rang et qu'on n'observe qu'un rang par observation.

obs.	X_1	X_2	*co-cluster*
u_1		1	$c_{.,0}$
...		...	...
u_{35}		35	$c_{.,0}$
u_{36}		36	$c_{.,1}$
...		...	...
u_{65}		65	$c_{.,1}$
u_{66}		66	$c_{.,2}$
...		...	...
u_{100}		100	$c_{.,2}$

$$\frac{1}{35!} = \frac{1}{1{,}03.10^{40}} \;\; ; \;\; \frac{1}{30!} = \frac{1}{2{,}65.10^{32}} \;\; ; \;\; \frac{1}{35!} = \frac{1}{1{,}03.10^{40}}$$

On choisit la configuration ci-contre parmi les $1{,}03.10^{40} \times 2{,}65.10^{32} \times 1{,}03.10^{40} = 2{,}83.10^{112}$ configurations possibles.

Notons que pour chaque terme de la vraisemblance, la probabilité que des données soient engendrées, alors qu'elles ne respectent pas les paramètres spécifiée dans la loi a priori, est nulle. Le produit de l'a priori et de la vraisemblance

du modèle donne la probabilité a posteriori du modèle connaissant les données. Le critère optimisé par l'approche MODL est le logarithme négatif de cette probabilité a posteriori.

Définition 15. *Un modèle de co-clustering $\mathcal{M}$ est optimal s'il minimise le critère ξ :*

$$
\begin{aligned}
\xi(\mathcal{M}) &= -\log P(\mathcal{M}) - \log \mathcal{L}(\mathcal{M}) \\
&= \sum_{j=1}^{d} \left(\mathbb{1}_{nom.}(X_j) \log n_j + \mathbb{1}_{cont.}(X_j) \log m \right) \\
&+ \sum_{j=1}^{d} \mathbb{1}_{nom.}(X_j) \log B(n_j, k_j) \\
&+ \log \binom{m+k-1}{k-1} \\
&+ \sum_{j=1}^{d} \mathbb{1}_{nom.}(X_j) \sum_{i_j=1}^{k_j} \log \binom{m^C_{\dots i_j \dots} + n^C_{\dots i_j \dots} - 1}{n^C_{\dots i_j \dots} - 1} \\
&+ \log m! - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \dots \sum_{i_d=1}^{k_d} \log m^C_{i_1 i_2 \dots i_d}! \\
&+ \sum_{j=1}^{d} \mathbb{1}_{nom.}(X_j) \left(\sum_{i_j=1}^{k_j} \log m^C_{\dots i_j \dots}! - \sum_{i_j=1}^{n_j} \log m_{\dots i_j \dots}! \right) \\
&+ \sum_{j=1}^{d} \mathbb{1}_{cont.}(X_j) \sum_{i_j=1}^{k_j} \log m^C_{\dots i_j \dots}!
\end{aligned}
\tag{2.16}
$$

où $\mathbb{1}_{cont.}(X_j)$ et $\mathbb{1}_{nom.}(X_j)$ indiquent le type de la variable X_j, respectivement continue et nominale.

Les quatre premières lignes du critère correspondent au logarithme négatif de la probabilité a priori du modèle et les trois dernières au logarithme négatif de la vraisemblance. En théorie de l'information, un logarithme négatif de probabilité est équivalent à la longueur de codage de Shannon-Fano Shannon1948. Le logarithme négatif de la probabilité a priori du modèle $-\log P(\mathcal{M})$ correspond donc à la longueur de codage du modèle de coclustering et le logarithme négatif de la vraisemblance $-\log \mathcal{L}(\mathcal{M})$ à la longueur de description des données. Minimiser la somme des deux revient alors à chercher le modèle de longueur et de description minimale, le critère a donc une interprétation en terme de *MDL (Minimum Description Length)*.

D'un point de vue algorithmique, l'optimisation est réalisée à l'aide d'une heuristique gloutonne ascendante, initialisée avec le modèle le plus fin, c'est-à-dire avec une modalité par cluster pour chacune des d variables. Elle considère

toutes les fusions entre les clusters et réalise la meilleure d'entre elles si cette dernière permet de faire décroître le critère. Cette heuristique est améliorée avec une étape de post-optimisation, pendant laquelle on effectue des permutations au sein des clusters. Le tout est englobé dans une méta-heuristique de type VNS (Variable Neighborhood Search, Hansen et Mladenovic (2001)) qui tire profit de plusieurs lancements de l'algorithme avec des initialisations aléatoires différentes. L'algorithme est détaillé et évalué par Boullé (2008). La complexité algorithmique est en $O(m\sqrt{m}\log(m))$. Cette complexité est calculée dans le pire des cas, c'est-à-dire lorsque la matrice des données est pleine. En pratique, l'algorithme est capable d'exploiter l'aspect creux habituellement observé dans ce type de données.

MODL est une approche de coclustering en d dimensions pouvant traiter des variables continues et nominales. De ce fait, elle est adaptée à de nombreux problèmes comme l'analyse de graphes ou de courbes. Ces applications sont détaillées dans le chapitre 3.

Boullé (2012, EGC) a démontré que le critère MODL dans le cas d'un bi-clustering de variables nominales converge asymptotiquement vers l'information mutuelle entre les variables X_1 et X_2. Dhillon *et al.* (2003) minimisent la perte de l'information mutuelle liée à la partition des valeurs des variables X_1 et X_2. Cette approche est équivalente à maximiser l'information mutuelle entre X_1^M et X_2^M puisque l'information mutuelle entre deux variables descriptives des données est un terme constant. L'approche MODL est donc asymptotiquement assimilable à l'approche de Dhillon *et al.* (2003). Cependant MODL est régularisée, ce qui en fait une approche plus fiable en régime non-asymptotique. D'autre part, l'approche MODL ne requiert pas de paramétrisation de la part de l'utilisateur, ce qui la rend plus simple à utiliser.

2.3 Bilan

Dans ce chapitre, nous avons, dans un premier temps, introduit les notions clés des approches de clustering. Dans un second temps, nous avons passé en revue les principales approches de co-clustering. Il apparaît que la majorité des problèmes identifiés pour le clustering est rencontrée dans de nombreuses approches de co-clustering. On peut noter par exemple le choix du nombre de clusters qui est un problème récurrent dans les méthodes de co-clustering. Le choix d'une dissimilarité est également nécessaire dans plusieurs approches, notamment les approches hiérarchiques.

L'approche MODL permet de traiter des données nominales ou continues, avec deux ou plusieurs variables, et ce, sans requérir de l'utilisateur le moindre paramètre de modélisation. Nous décidons donc de choisir cette approche comme base pour le reste de la thèse.

Les problèmes de l'évaluation et de l'exploitation du co-clustering sont plus rarement traités. Nous proposons, dans le reste de la thèse, de modéliser

certains problèmes de data mining à l'aide du co-clustering. La pertinence de l'utilisation d'une telle approche sur les problèmes traités est évaluée grâce à un protocole expérimental sur des données artificielles. On a vu que l'approche MODL possède des propriétés algorithmiques permettant de réaliser des co-clustering sur des données volumétriques. Nous verrons que, dans ce cas, il est important de savoir extraire une information pertinente des résultats.

3

Le co-clustering, applications diverses

Le co-clustering est une technique permettant de réaliser des classifications croisées de données décrites par des variables de type numérique ou nominal. Cette technique peut donc être appliquée à de nombreux types de problèmes de classification non-supervisée. Ce chapitre a pour but de donner un panorama des différentes utilisations possibles du co-clustering, depuis le cas classique du co-clustering de deux variables nominales au cas général à d variables hétérogènes.

3.1 Objectifs et contributions

Dans ce chapitre, nous présentons des utilisations possibles du co-clustering avec deux dimensions ou plus. Les dimensions des données sont les modalités des variables, qui sont simultanément partitionnées par l'approche MODL. Pour chacune d'elles, nous suivons la même démarche :

- un positionnement du problème,
- la formalisation du problème de co-clustering,
- des expérimentations sur données artificielles,
- des expérimentations comparatives.

La première étude concerne le partitionnement de graphe. La modélisation de ce problème avec l'approche MODL a déjà été proposée par Boullé (2011). Nous proposons dans ce chapitre un positionnement différent, orienté réseaux sociaux, et effectuons des expérimentations comparatives sur des données engendrées selon deux modèles différents. Cette étude nous sert de base pour proposer une modélisation adaptée à l'analyse des graphes temporels. Après avoir réalisé un état de l'art, nous proposons une modélisation du problème permettant une segmentation simultanée des nœuds et du temps. Des expérimentations sur des données artificielles montrent la pertinence de la modélisation : l'approche est testée sur des graphes aléatoires, stationnaires (c'est-à-dire sans évolution temporelle) et sur des graphes dont nous connaissons la structure et l'évolution temporelle sous-jacente. Nous ne réalisons pas d'expérimentations comparatives, faute d'implémentations disponibles d'approches alternatives. Enfin, nous nous intéressons au problème du clustering de courbes, formalisé et étudié par Boullé (2012). Dans cette étude, nous utilisons une variable nominale « fictive » correspondant aux identifiants des courbes. Nous proposons dans ce chapitre un positionnement ainsi que des analyses sur des données artificielles, engendrées suivant quatre motifs différents.

3.2 Le biclustering nominal pour le partitionnement de graphes

Dans un premier temps, nous étudions le cas le plus simple de co-clustering, à savoir le biclustering ou co-clustering en deux dimensions. Cette technique a déjà été largement étudiée, notamment dans le domaine de la bio-statistique ou du traitement des données textuelles. Une autre application est l'étude de graphes. Nous proposons, dans cette section, de présenter les différentes approches de partitionnement de graphes, de voir en quoi le biclustering est particulièrement adapté à ce type de problèmes, d'introduire l'approche MODL pour les graphes et son utilisation pour la détection de structures dans les graphes.

3.2.1 Rappels de théorie des graphes

Rappelons la terminologie et les principales définitions de théorie des graphes. Un graphe est un ensemble de *nœuds* (ou *sommets*) reliés entre eux par des *liens*. Il existe plusieurs types de graphes :

- les graphes non-orientés : les nœuds sont reliés entre eux par des liens symétriques appelés *arêtes*,
- les graphes orientés : les liens (appelés ici *arcs*) ont un nœud de départ (la *source*) et un nœud d'arrivée (la *cible*),
- les multigraphes : plusieurs liens (arcs ou arêtes) peuvent lier deux mêmes nœuds du graphe,
- les graphes bipartis : le graphe possède deux ensembles de nœuds distincts correspondant à la source et à la cible des liens.

Ainsi, les graphes sont des représentations appropriées pour la modélisation de réseaux ou de données relationnelles. On peut, par exemple, représenter un réseau social par un graphe simple (graphe non-orienté sans liens multiples et sans boucles) où les nœuds modélisent des *acteurs* et les arêtes le lien qui les unissent. On peut également utiliser un multigraphe orienté pour modéliser un flux de personnes dans un réseau de transports. Enfin, un multigraphe biparti peut être utilisé pour modéliser un ensemble de transactions de type achats de produits par des clients.

La notion de *degré* est importante pour la suite de l'étude. Il s'agit du nombre d'arcs sortant d'un nœud source ou entrant dans un nœud cible. Dans un graphe de transaction par exemple, le degré d'un nœud de type « client » est le nombre total d'achats qu'il a effectués.

3.2.2 État de l'art des approches de clustering de graphes

Le clustering de graphe a été largement étudié. De nombreux états de l'art présentent un large panorama des méthodes existantes (Schaeffer, 2007; Goldenberg *et al.*, 2009; Fortunato, 2010), parmi lesquelles on peut compter les clusterings hiérarchiques basés sur une mesure de similarité entre les nœuds (Hastie *et al.*, 2009) ou encore sur les méthodes spectrales (Pothen *et al.*, 1990). Ici, les approches de type *Blockmodeling* (ou modélisation par blocs) pour la segmentation de graphes sont mises en avant. Ce choix est motivé par la nature commune des structures recherchées par ces méthodes et par l'approche MODL.

Des premières analyses de réseaux sociaux à l'introduction du blockmodeling Dès le milieu des années 1950, Nadel (1957) introduit la notion de *rôle* pour décrire le comportement d'acteurs au sein d'un réseau et s'intéresse à leurs interrelations. Il part du principe selon lequel un réseau possède une structure sociale s'il existe des motifs caractéristiques dans les relations entre les acteurs du réseau.

Ces travaux sont repris par Lorrain et White (1971) qui définissent la notion d'*équivalence structurelle* : dans un réseau, deux individus jouent le même rôle s'ils interagissent de la même manière avec le reste du graphe. Ainsi, les acteurs sont regroupés afin de construire une représentation synthétique de la structure du graphe. Les classes définies sont décrites par le rôle qu'elles ont vis-à-vis de l'ensemble des individus du réseau.

White *et al.* (1976) proposent une représentation matricielle du graphe où les éléments de la matrice modélisent les interactions entre les nœuds représentés par les lignes et les colonnes. Ce type de représentation est par la suite désigné par le terme *matrice d'adjacence* (voir exemple figure 3.1.c). Ils proposent de regrouper les individus jouant un même rôle en ordonnant les lignes et les colonnes de la matrice d'adjacence afin de partitionner cette dernière en *blocs*. On peut ainsi résumer une matrice de taille n^2, avec n le nombre d'acteurs, par une matrice de taille k^2, avec k le nombre de groupes de lignes et colonnes. Cette matrice résumée est la matrice d'adjacence d'un nouveau graphe, dit *graphe image* (voir exemple figure 3.1.b), qui est une représentation plus interprétable du graphe initial. Cette conception de la modélisation des rôles dans un réseau social est nommée *blockmodeling*.

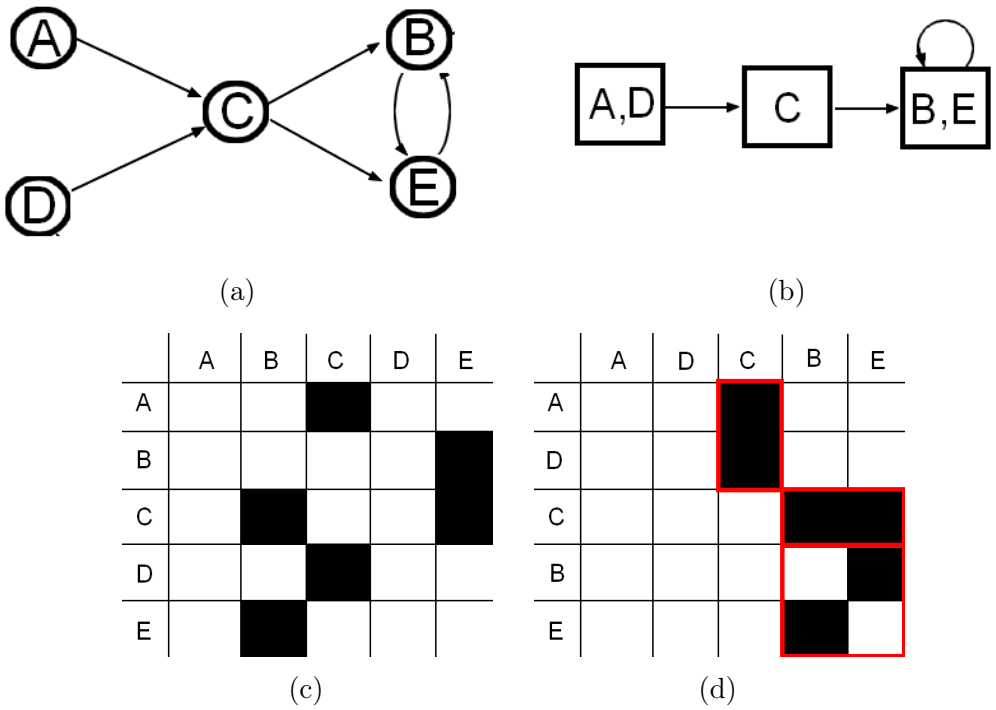


FIGURE 3.1 – (a) exemple d'un graphe orienté constitué de 5 acteurs. (c) est la matrice d'adjacence correspondant à ce graphe. Le graphe (b) est le graphe image de (a) après avoir affecté 3 rôles aux acteurs de celui-ci. (d) est la représentation sous forme de blocs de la matrice (b), les colonnes et les lignes ont été réordonnées pour mettre en relief la structure sous-jacente du graphe.

White et Reitz (1983) introduisent la notion d'*équivalence régulière* comme une généralisation de l'équivalence structurelle. À la différence du cas structurel qui requiert que des acteurs ayant des rôles similaires interagissent avec les mêmes acteurs, l'équivalence régulière se veut plus souple en nécessitant que deux acteurs interagissent de manière identique avec les mêmes clusters pour être groupés ensemble. Ainsi, dans le cas de la figure 3.1, les nœuds A et D sont équivalents structurellement et régulièrement, et les nœuds B et E ne sont pas équivalents structurellement mais le sont régulièrement. Dans le cas des cliques, c'est-à-dire lorsque les acteurs sont tous connectés les uns aux autres, les acteurs sont régulièrement équivalents mais pas structurellement.

De nombreuses approches de partitionnement de graphes sont basées sur le blockmodeling. Elles se décomposent en plusieurs sous familles.

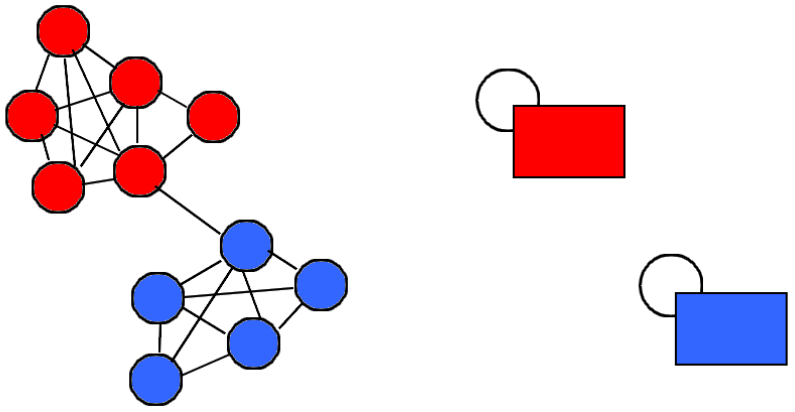
Le blockmodeling déterministe. Les premiers algorithmes proposés pour extraire des blocs dans les matrices d'adjacence se sont appuyés sur les définitions de l'équivalence structurelle. Avec l'algorithme *CONCOR*, Breiger *et al.* (1975) proposent une classification hiérarchique descendante basée sur l'étude des corrélations entre les lignes et les colonnes de la matrice.

Suite à l'introduction de la notion d'équivalence régulière, les techniques ont évolué. Batagelj *et al.* (1992) proposent l'optimisation d'une fonction objectif pour isoler les blocs vides. D'abord utilisée sur les graphes simples, cette technique est étendue aux multigraphes bipartis par Doreian *et al.* (2004).

La principale direction prise dans le blockmodeling déterministe s'oriente vers l'optimisation de critères évaluant la pertinence du graphe image en tant que représentation synthétique du graphe d'origine. Wasserman et Faust (1994) proposent un ensemble de mesures permettant de quantifier la qualité du schéma en blocs du graphe.

La maximisation de modularité et la méthode de Clique Percolation. Une des approches les plus populaires en analyse de réseaux sociaux consiste à maximiser la modularité (Blondel *et al.*, 2008; Rossi et Villa-Vialaneix, 2012). La modularité, introduite par Girvan et Newman (2002), est une mesure définie sur $[-1, 1]$ qui compare le nombre d'arêtes dans les clusters avec le nombre d'arêtes observées dans un graphe aléatoire de même taille et avec la même répartition des degrés. En maximisant cette mesure, on obtient des clusters de nœuds fortement connectés entre eux et faiblement connectés au reste du graphe, appelés aussi *Communautés*. En outre, on se place dans le cas très particulier d'un blockmodeling diagonal avec correction des degrés (Reichardt et White, 2007).

Comme il a été vu précédemment, de nombreux algorithmes de blockmodeling cherchent à faire en sorte qu'un graphe image défini par certains paramètres, se calque sur le graphe d'origine. Wasserman et Faust (1994) proposent des mesures de qualité qui permettent de juger de la cohérence du graphe image par rapport au graphe initial. Ici, la modularité est également une mesure de



(a) (b)

FIGURE 3.2 – (a) Graphe non-pondéré contenant deux sous-ensemble de nœuds fortement connectés entre eux. (b) représentation du graphe image correspondant.

qualité qui détermine si un graphe image, dont l'ensemble des arêtes sont des boucles, représente bien le graphe étudié, comme c'est le cas dans la figure 3.2.

La maximisation de modularité ne respecte pas les assertions de Lorrain et White (1971) dans le sens où elle ne considère que les interactions internes aux clusters. De plus, malgré un sens statistique du critère, Bickel et Chen (2009) ont démontré que son optimisation est asymptotiquement biaisée et peut conduire à la construction de structures incorrectes. Néanmoins, bien que la maximisation de modularité soit un problème NP difficile (Brandes *et al.*, 2006), des algorithmes (Blondel *et al.*, 2008; Rossi et Villa-Vialaneix, 2012) permettent d'obtenir une partition satisfaisante des graphes dans des temps raisonnables, ce qui en fait une technique populaire dans la littérature.

Le blockmodeling stochastique. Holland *et al.* (1983) définissent la notion d'équivalence stochastique : deux nœuds sont stochastiquement équivalents si leurs lois de probabilités de connexions avec les autres acteurs du réseau sont les mêmes.

Nowicki et Snijders (2001) proposent un algorithme de blockmodeling stochastique avec un nombre de clusters fixe. À chaque nœud du graphe est associé une variable latente z tirée suivant une loi multinomiale de paramètre $(1, \alpha)$, un vecteur de taille k le nombre de clusters, dont les éléments sont les proportions de chaque cluster. La variable z est également un vecteur de taille k dont l'élément modélisant l'appartenance à un cluster est égal à 1, les autres à 0. Les arcs sont finalement tirés conditionnellement à la variable latente suivant une loi de Bernoulli prenant en paramètre la matrice des probabilités de connexion entre deux clusters. La loi de probabilité *a posteriori* des paramètre du modèle est approchée par un échantillonnage de Gibbs.

Kemp et Tenenbaum (2006) proposent d'étendre le précédent algorithme de blockmodeling stochastique en définissant un modèle relationnel infini. Ce modèle traite une matrice binaire exprimant les relations entre nœuds source et nœuds cible. Ce modèle est infini car le nombre de clusters est potentiellement infini en régime asymptotique, c'est-à-dire lorsque le nombre d'arcs et de nœuds tendent vers l'infinis. La principale différence avec l'approche de Nowicki et Snijders (2001) réside dans la probabilité a priori des paramètres de modélisation, probabilité sur la variable latente qui permet de déterminer le nombre de clusters. Ce nombre et l'affectation des éléments dans les clusters sont déterminés à l'aide d'un processus du restaurant chinois (Pitman, 2006) de paramètre de concentration γ. L'idée est d'initialiser le clustering avec un seul cluster contenant un unique nœud, puis d'ajouter des éléments un par un, chacun des clusters attirant des nouveaux nœuds en fonction de son volume et du paramètre de concentration γ. Ainsi à chaque itération, soit le nœud est placé dans un cluster, soit il permet la création d'un nouveau cluster dont il est l'unique élément. À chaque ajout d'élément, la probabilité d'observer des liens entre les clusters est spécifiée et suit une loi Beta de paramètre β. Enfin, la probabilité d'observer des liens entre les nœuds suit une loi de Bernoulli définie suivant les paramètres de précédents.

Airoldi *et al.* (2008) proposent le *mixed-membership stochastic blockmodeling*. Contrairement aux approches précédentes, la variable latente n'est pas un vecteur contenant un 1 associant un nœud à un cluster, et des 0 ailleurs ; mais un ensemble de réels compris entre 0 et 1 et sommant à 1, indiquant le degré d'appartenance d'un nœud à l'ensemble des clusters. L'ensemble des paramètres de modélisation est inféré par des méthodes variationnelles.

Les graphes bipartis. Dans certains cas, il peut y avoir deux types d'acteurs dans un graphe, ainsi la matrice d'adjacence n'est plus carrée et le schéma de clustering est différent sur les lignes et les colonnes. C'est le cas, par exemple, des graphes représentant les achats de produits par des clients, les goûts musicaux, voire des graphes de co-citations ou d'envois de mails où il peut être intéressant d'avoir des clusters d'acteurs source et d'acteurs cible. Dhillon *et al.* (2003) appliquent une méthode de co-clustering sur une table de contingence normalisée traitée comme une loi de probabilité jointe entre deux variables nominales prenant leurs valeurs sur les lignes et les colonnes de la matrice. Les biclusters sont obtenus par minimisation de la perte d'information mutuelle entre les variables segmentées en clusters et les données initiales. Cette méthode, illustrée par des exemples de coclustering textes/mots, peut être étendue aux graphes bipartis comme suggéré par Karrer et Newman (2010). Cependant, bien que le critère puisse être optimisé de manière efficace avec des heuristiques, le nombre de clusters doit être spécifié. Il s'agit d'un paramètre qui peut être difficile à déterminer, notamment dans le cas de graphes de grande taille.

3.2.3 MODL pour le partitionnement de graphes

L'approche MODL se prête bien au partitionnement de graphes. Différents types de graphes peuvent être étudiés. Le cas le plus général est celui des multigraphes bipartis orientés, c'est-à-dire le cas où la matrice d'adjacence est rectangle (donc asymétrique) avec le nombres d'arcs dans les cases de la matrice. Mais les autres types de graphes peuvent également être traités. Par exemple, un graphe non orienté est représenté par une matrice d'adjacence carrée et symétrique et peut être traité de la même façon qu'un graphe biparti.

Dans le cas des graphes, les variables étudiées sont les nœuds source et les nœuds cible. L'unité statistique est l'arc : chaque arc est une observation décrite par son nœud source et son nœud cible. On a donc affaire à un biclustering (ou co-clustering en deux dimensions) des valeurs de variables nominales. Le tableau 3.2 donne les notations.

Le graphe $\mathcal{D}$	**Le graphe image $\mathcal{M}$**
X_1 les nœuds source n_1 le nombre de nœuds source $m_{i_1.}$ degré du nœud source i_1	X_1^M la partition des nœuds source k_1 le nombre de clusters source $m^C_{i_1.}$ degré du cluster source i_1
X_2 les nœuds cible n_2 le nombre de nœuds cible $m_{.i_2}$ degré du nœud cible i_2	X_2^M la partition des nœuds cible k_2 le nombre de clusters cible $m^C_{.i_2}$ degré du cluster cible i_2
m le nombre d'arcs	$k = k_1 k_2$ le nombre de biclusters $m^C_{i_1 i_2}$ effectif du bicluster formé par les clusters i_1 et i_2

TABLE 3.1 – Notations pour les graphes

Comme il a été vu dans le chapitre 2.2.5, l'approche MODL infère à partir des données de graphe $\mathcal{D}$, le meilleur modèle $\mathcal{M}$, ici le graphe image. Le graphe image est défini par un ensemble de paramètres, sur lesquels sont faites des hypothèses de loi a priori uniforme :

1. **les nombres de clusters de nœuds**. Il s'agit ici de fixer les nombres k_1 et k_2 de groupes qu'on souhaite obtenir. Le choix des nombres de clusters est réalisé indépendamment pour les deux variables. A priori, ces nombres sont inconnus. On considère donc qu'ils sont compris entre un et le nombre de nœuds. Lorsque le graphe n'est pas orienté et non-biparti, le nombre de clusters de nœuds source et de nœuds cible sont identiques. Le cas avec un unique cluster correspond à un graphe image où aucune structure significative n'est capturée. Le cas avec autant de clusters que de nœuds modélise le graphe image le plus fin, où chacun des acteurs du réseau joue un rôle suffisamment caractéristique pour qu'il soit différencié des autres. Ces deux schémas sont en accord avec la définition de l'équivalence régulière (White et Reitz, 1983; Borgatti, 1988) ;

2. **la répartition des nœuds dans les clusters**. Une fois les nombres de clusters k_1 et k_2 connus, on choisit une partition uniformément parmi l'ensemble des partitions possibles. Le choix des partitions est réalisé indépendamment pour les deux variables. Il est difficile de savoir comment se répartissent les nœuds au sein des clusters. Le choix est donc fait de considérer équiprobables toutes les partitions de n_1 (resp. n_2) nœuds dans k_1 (resp. k_2) clusters, potentiellement vides. Le nombre de répartitions possibles $B(n_1, k_1)$ (resp. $B(n_2, k_2)$) équivaut à une somme de nombres de Stirling de second ordre. En comparaison, Kemp et Tenenbaum (2006) utilisent un processus du restaurant chinois pour choisir à la fois le nombre de clusters et leurs effectifs. Les répartitions avec un gros cluster et de nombreux petits clusters sont ici favorisées et donc les structures avec des clusters équilibrés demandent un plus grand nombre d'observations pour être détectées. Les hypothèses faites dans l'a priori de l'approche MODL ne favorisent pas une répartition en particulier, comme c'est le cas dans l'approche de Kemp et Tenenbaum (2006).
3. **la répartition des arcs dans les biclusters**. Le nombre de clusters de chaque partition des nœuds source et cible permet de connaître le nombre de biclusters $k = k_1 k_2$. On connaît également le nombre total d'arcs m du graphe. On peut donc spécifier le nombre d'arcs reliant les clusters. On considère ici aussi que toutes les configurations sont équiprobables ;
4. **la répartition du degré des clusters sur leurs nœuds**. Le degré des clusters est déduit des effectifs des arcs reliant les clusters, spécifiés précédemment. Le nombre de nœuds de chaque cluster est également connu. On considère équiprobables, pour chaque cluster, toutes les façons de distribuer le degré du cluster source $m^C_{i_1.}$ (resp. cluster cible $m^C_{.i_2}$) sur les $n_{i1.}$ (resp. $n_{.i_2}$) nœuds qu'il contient. La répartition des degrés est réalisée indépendamment pour chacun des clusters.

Ces paramètres spécifiés, le critère décrit en section 2.2.5 peut se simplifier comme décrit dans l'équation 3.1.

Définition 16. *Le meilleur modèle $\mathcal{M}$ (ou graphe image dans le cas des graphes) est obtenu en maximisant le critère MODL ξ.*

$$\begin{aligned}\xi(\mathcal{M}) = &\log n_1 + \log n_2 \\ &+ \log B(n_1, k_1) + \log B(n_2, k_2) \\ &+ \log \binom{m+k-1}{k-1} \\ &+ \sum_{i_1=1}^{k_1} \log \binom{m^C_{i_1.} + n^C_{i_1.} - 1}{n^C_{i_1.} - 1} + \sum_{i_2=1}^{k_2} \log \binom{m^C_{.i_2} + n^C_{.i_2} - 1}{n^C_{.i_2} - 1} \\ &+ \log m! - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \log m^C_{i_1 i_2}! \\ &+ \sum_{i_1=1}^{k_1} \log m^C_{i1.}! - \sum_{i_1=1}^{n_1} \log m_{i1.}! + \sum_{i_2=1}^{k_2} \log m^C_{.i2}! - \sum_{i_2=1}^{n_2} \log m_{.i2}!\end{aligned} \tag{3.1}$$

Les quatre premières lignes du critère correspondent aux termes de l'a priori : la première spécifie le nombre de clusters pour chacun des ensembles de nœuds source et cible, la seconde la partition des nœuds en clusters, la troisième la répartition des arcs dans les biclusters et la quatrième la répartition du degré des clusters sur les nœuds qu'ils contiennent. Les deux dernières lignes sont les termes de vraisemblance développés dans le chapitre 2.2.5 et adaptés au cas d'un biclustering de deux variables nominales.

3.2.4 Expérimentations sur des multigraphes non-orientés

Dans cette partie, nous construisons des graphes artificiels non-orientés et à arcs multiples. Nous utilisons des graphes avec une structure en quasi-cliques. Un graphe composé de quasi-cliques est un graphe qui peut se partitionner en sous-graphes dont les nœuds sont fortement connectés entre eux et très peu connectés avec les nœuds des autres sous-graphes. Cette structure de graphe est recherchée par les approches de clustering basées sur la maximisation de modularité. Nous comparons l'approche MODL avec la maximisation de modularité (Girvan et Newman, 2002) sur les graphes artificiels.

Construction des graphes On considère un multigraphe non-orienté. Les variables X_1 et X_2, ainsi que les variables-partitions X_1^M et X_2^M sont identiques. Afin d'alléger la notation, on note X la variable prenant des valeurs sur l'ensemble des n nœuds et X^M la variable-partition prenant des valeurs sur l'ensemble des k clusters. L'ensemble des m arcs est nommé E. Θ est une matrice de taille n^2 dont les éléments sont les probabilités de connexion de chaque paire de nœuds. On note p la proportion d'arêtes observées à l'intérieur

des clusters et $1-p$ la proportion d'arêtes observées entre les clusters, avec $p > 0.5$. Les graphes sont engendrés suivant le protocole suivant :

- choix du nombre de clusters : $k = 10$;
- choix du nombre de nœuds : $n = 100$;
- partition des n nœuds en k clusters : $X^M \sim multinomiale\left(n; \frac{1}{k}, ..., \frac{1}{k}\right)$;
- construction de la matrice triangulaire supérieure de probabilités de connexion des n nœuds :

$$\Theta = \begin{cases} \theta_{i_1 i_2} = p & \text{si } i_1 \leq i_2 \text{ et } x_{i_1}, x_{i_2} \text{ sont dans le même cluster,} \\ \theta_{i_1 i_2} = 1-p & \text{si } i_1 \leq i_2 \text{ et } x_{i_1}, x_{i_2} \text{ ne sont pas dans le même cluster,} \\ \theta_{i_1 i_2} = 0 & \text{sinon ;} \end{cases}$$

- la moitié des arcs est tirée sur U la matrice triangulaire supérieure de la matrice d'adjacence :

$$U \sim \ multinomiale\left(\frac{m}{2}, \frac{\Theta}{\sum_{i_1,i_2} \theta_{i_1 i_2}}\right) \quad i_1 = 1..n \quad i_2 = 1..n$$

- on construit A la matrice d'adjacence symétrique pour que le graphe soit non-orienté : $A = \{a_{i_1 i_2} = u_{i_1 i_2} + u_{i_2 i_1}\} \quad i_1 = 1..n \quad i_2 = 1..n$

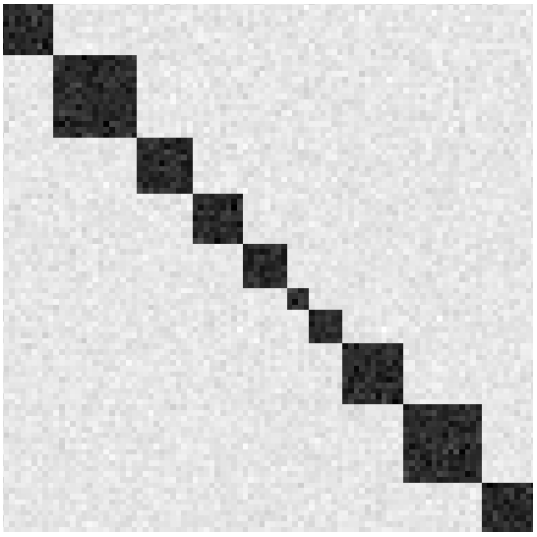


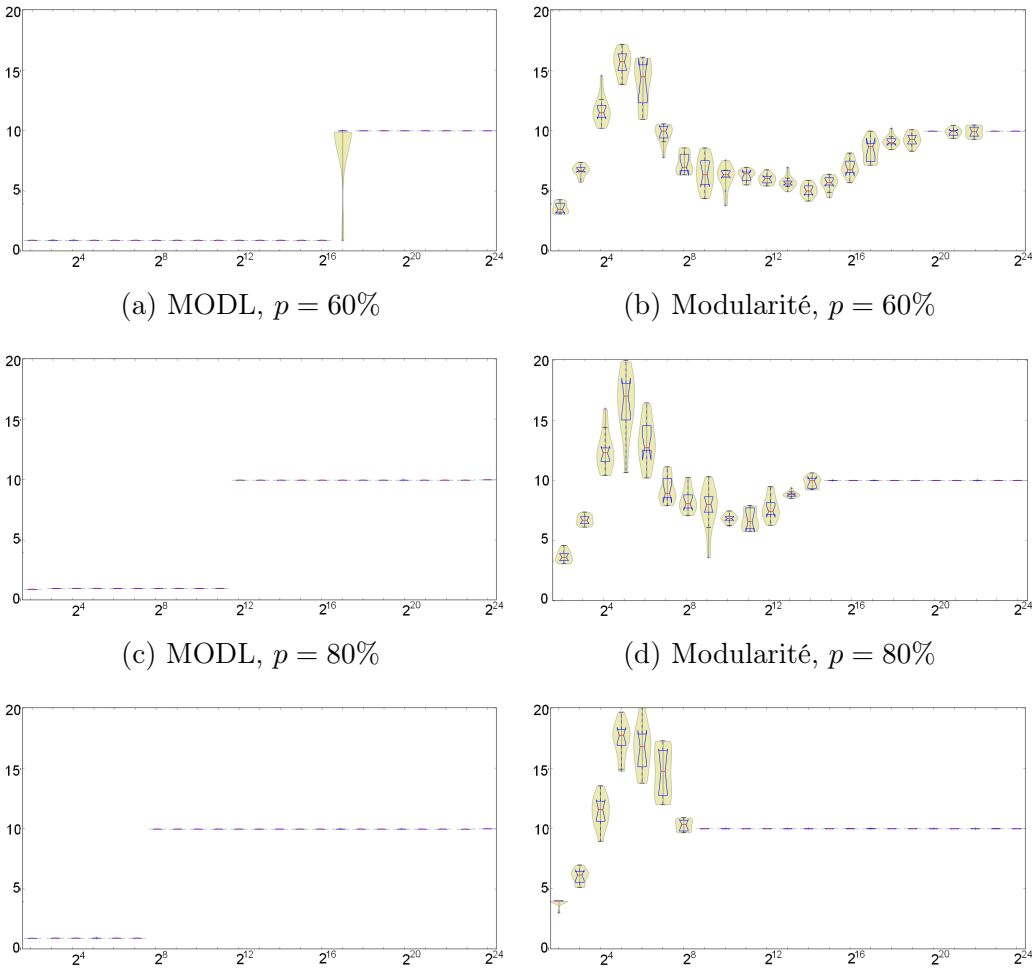
FIGURE 3.3 – Exemple de graphe modulaire avec $n = 100$ nœuds, $k = 10$ clusters, $m = 10^6$ arcs et $p = 0.8$ la proportion d'arêtes observées à l'intérieur des clusters.

Les expérimentations ont été effectuées avec un nombre croissant d'arêtes m et avec des valeurs différentes de p. Pour la construction de l'intégralité des graphes, les clusters de nœuds demeurent inchangés. Pour chaque combinaison de m et p, dix graphes différents sont engendrés. Dans ce protocole expérimental, nous engendrons des graphes avec des structures en co-cliques avec un croissant d'arêtes. Dans l'approche MODL, les observations sont les arcs. Nos données sont engendrées suivant un modèle génératif proche de celui de MODL mais avec structure contrainte, favorable aux approches de maximisation de modularité.

Expérimentations Nous comparons deux approches de partitionnement de graphe dans cette partie : une approche de maximisation de modularité adaptée aux multigraphes et l'approche MODL. La manière dont sont engendrées les données est adaptée aux deux approches. La maximisation de modularité cherche à segmenter le graphe en cliques, alors que l'approche MODL cherche des co-clusters d'arcs de densités homogènes. L'approche MODL est implémentée dans le logiciel *Khiops,* qui a été utilisé pour cette étude. Pour la maximisation de modularité, nous utilisons l'implémentation de Rossi et Villa-Vialaneix (2012). Cette implémentation compare la valeur de la modularité obtenue pour la meilleure partition du graphe étudié à la valeur obtenue en partitionnant des graphes aléatoires dont la répartition des degrés est identique à la répartition des degrés du graphe étudié. Ce correctif permet à la maximisation de modularité d'être plus robuste vis à vis des graphes aléatoires. Afin de s'assurer de cette propriété, nous avons cherché à partitionner cent graphes d'Erdős-Renyi de tailles différentes. Il s'agit de graphes aléatoires, construits de manière à ce que la probabilité de connexion des nœuds soit la même pour chaque couple de nœuds du graphe. L'implémentation de la maximisation de modularité n'a jamais segmenté les graphes. Les résultats obtenus sur les mêmes graphes avec l'approche MODL sont identiques.

Résultats Les graphiques de la figure 3.4 et 3.5 sont appelés diagrammes à violons (Hintze et Nelson, 1998). Il s'agit de boites à moustaches, combinées avec des estimateurs de densité locaux à chacune des séries de mesures. Cette visualisation nous permet d'étudier les statistiques de bases comme la médiane, les quartiles ainsi que les valeurs minimales et maximales. De plus l'estimateur de densité nous donne une idée de la répartition des points de la série des mesures d'ARI pour chaque paramètre des graphes engendrés. Dans notre étude, nous traçons un diagramme pour chaque ensemble de graphes, construits avec les même paramètres.

L'évolution du nombre de clusters retrouvés par l'approche MODL et par l'approche de maximisation de modularité, en fonction du nombre d'arêtes engendrées, est tracée dans la figure 3.4. On observe que pour un faible nombre d'arêtes, l'approche MODL ne retrouve aucune structure : l'information présente dans les données ne permet pas de faire émerger de partition du graphe. Suit une période transitoire très brève (observée seulement pour $p = 0{,}6$) où,



(a) MODL, $p = 60\%$ (b) Modularité, $p = 60\%$

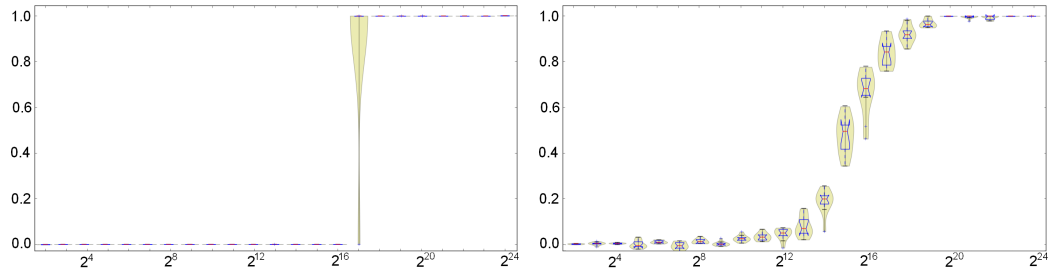
(c) MODL, $p = 80\%$ (d) Modularité, $p = 80\%$

(e) MODL, $p = 100\%$ (f) Modularité, $p = 100\%$

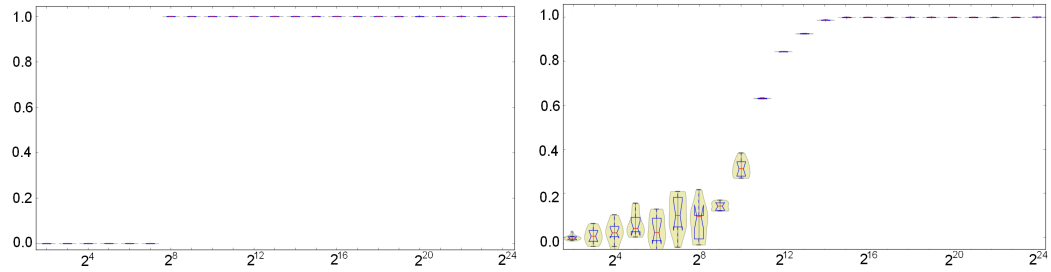
FIGURE 3.4 – Nombre de clusters trouvés en fonction du nombre d'arêtes engendrées pour des proportions d'arêtes dans les co-clusters diagonaux de 60%, 80% et 100%.

pour certains graphes, les arêtes sont en quantité suffisante pour faire émerger la structure engendrée, alors que pour d'autres, les données ne sont pas en quantité suffisante pour que le graphe soit segmenté. Enfin, le régime asymptotique est atteint : le nombre de clusters engendrés est retrouvé et l'ARI est maximal (voir figure 3.5), ce qui prouve que la structure trouvée est la structure engendrée artificiellement. Le régime asymptotique est observé pour un nombre d'arêtes engendrées d'autant plus petit que la proportion d'arêtes dans les co-clusters diagonaux est proche de 1.

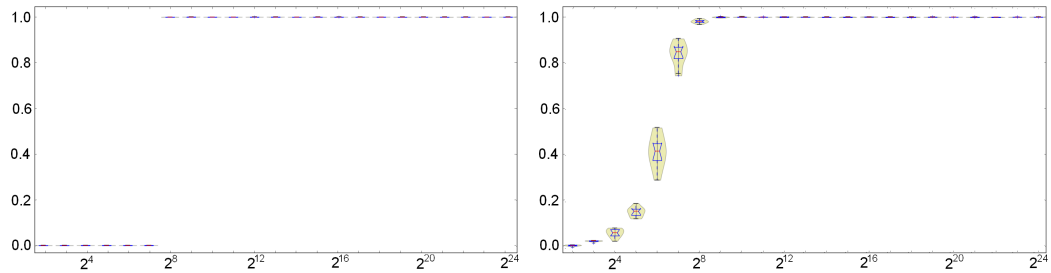
L'évolution de l'indice de Rand ajusté (ARI) entre la vraie partition du graphe et les partitions obtenues par l'approche MODL et par l'approche de maximisation de modularité, en fonction du nombre d'arêtes engendrées, est



(a) MODL, $p = 60\%$ (b) Modularité, $p = 60\%$



(c) MODL, $p = 80\%$ (d) Modularité, $p = 80\%$



(e) MODL, $p = 100\%$ (f) Modularité, $p = 100\%$

FIGURE 3.5 – Indice de Rand ajusté en fonction du nombre d'arêtes engendrées pour des proportions d'arêtes dans les co-clusters diagonaux de 60%, 80% et 100%.

tracée dans la figure 3.5. Contrairement aux résultats obtenus avec MODL, la maximisation de modularité produit systématiquement une partition des graphes. Pour un faible nombre d'arêtes engendrées (entre 2 et 2^8), la partition obtenue est composée de nombreux clusters (entre cinq et vingt). Bien que l'indice de Rand ajusté soit proche de zéro, ce résultat montre que les partitions trouvées diffèrent de la partition engendrée avec les paramètres précédemment fixés. À partir de 2^8 arêtes engendrées et pour une proportion d'arêtes inter-clusters nulle, la maximisation de modularité permet de retrouver la partition théorique engendrée. Pour une proportion d'arêtes inter-clusters non-nulle, la partition obtenue trouve une partie des clusters : l'ARI est positif et le nombre de clusters inférieur à dix, ce qui signifie que la structure en cliques est partiellement trouvée. Notons que les variances de l'ARI et du nombre de

clusters sont faibles, ce qui montre que la qualité des partitions trouvées est constante d'un graphe à l'autre. Enfin à partir d'un nombre d'arêtes engendrées (2^{20} lorsque la proportion d'arêtes dans les co-clusters diagonaux est de 60% et 2^{14} lorsqu'elle est de 80%), la structure théorique engendrée est retrouvée par la maximisation de la modularité. De la même manière que pour les partitions obtenues par l'approche MODL, le nombre d'arêtes nécessaires pour faire émerger la structure du graphe est d'autant plus important que la proportion d'arêtes inter-clusters est forte.

En termes d'ARI, la maximisation de modularité produit de meilleurs résultats que l'approche MODL car elle nécessite un nombre moins important d'arêtes pour faire apparaître une structure pertinente (ARI positif). Cependant, la variance des résultats est bien plus importante que la variance des résultats obtenus avec MODL. Cette stabilité des résultats montre la fiabilité de l'approche MODL. Notons que la variabilité des résultats obtenus par maximisation de la modularité peut être lié à de l'algorithme utilisé (Fortunato, 2010), dont les résultats dépendent de l'ordre de traitement des observations (ici les arêtes).

Pour un faible nombre d'arêtes, la maximisation de la modularité produit un très grand nombre de clusters (entre quinze et vingt) mais présente une ARI proche de zéro, cette partition fine est donc très différente de la partition théoriquement engendrée. Ce résultat est caractéristique des approches non-régularisées : la qualité de la partition est meilleure avec un grand nombre de clusters, mais le gain n'est pas substantiel par rapport à une partition parcimonieuse ou à l'absence de partition. Comme toute approche de clustering, on cherche à simplifier au maximum les données. On préfère donc une partition simple à une partition fine à qualités égales. Enfin, dernier point, la modularité est algorithmiquement plus rapide à optimiser. Les résultats ont été produits plus rapidement avec la maximisation de la modularité (environ deux minutes) qu'avec l'approche MODL (environ une heure).

Pour résumer, dans le cas où on sait a priori que le graphe peut se segmenter en cliques et que les arêtes sont en très grande quantité, l'utilisation de la maximisation de la modularité est privilégiée pour ses qualités algorithmiques. Sinon, on préfère l'approche MODL pour sa fiabilité et sa capacité à inférer des structures plus complexes, qui ne se limitent pas à des structures diagonales.

MODL et la maximisation de modularité, deux approches différentes. L'approche MODL fait l'hypothèse que les données sont engendrées suivant un modèle et cherche à inférer les paramètres de ce modèle alors qu'en maximisant la modularité, on cherche la meilleure segmentation des données observées. Les données de cette expérimentation ont été engendrées sous l'hypothèse qu'il existe un modèle génératif de données avec une structure en cliques. Dans le cas où les données (ici les arêtes) sont engendrées en faible quantité, la structure théorique du graphe n'est pas empiriquement observée, c'est-à-dire que les arcs

ne sont pas en nombre suffisant pour faire émerger les dix cliques du modèle génératif des données.

Pour illustrer ce phénomène, nous proposons d'étudier l'information mutuelle normalisée (NMI) entre les variables-partitions X_1^M et X_2^M.

Définition 17. *L'information mutuelle normalisée entre X_1^M et X_2^M est définie de la manière suivante (Cover et Thomas, 2006) :*

$$NMI(X_1^M; X_2^M) = \frac{2}{H(X_1^M) + H(X_2^M)} \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} P^C_{i_1 i_2} \log \frac{P^C_{i_1 i_2}}{P^C_{i_1 .} P^C_{. i_2}} \tag{3.2}$$

où H est l'entropie de Shannon (voir chapitre 4), $P^C_{i_1 i_2}$ est la probabilité d'observer des arcs dans le co-cluster (i_1, i_2), $P^C_{i_1 .}$ est la probabilité d'observer des arcs ayant le cluster i_1 pour origine et $P^C_{. i_2}$ est la probabilité d'observer des arcs ayant le cluster i_2 pour destination.

On calcule l'information mutuelle normalisée dans les cas suivants :

- *l'information mutuelle normalisée théorique (NMI théorique)* : l'information mutuelle normalisée entre les deux variables-partitions telles que la loi de probabilité jointe des variables-partitions soit la loi de probabilité jointe du modèle génératif des données, c'est-à-dire les proportions p et $1 - p$ normalisées. Cette valeur est constante quel que soit le nombre d'arcs engendrés ;
- *l'information mutuelle normalisée empirique (NMI empirique)* : l'information mutuelle normalisée entre les deux variables-partitions telles que la loi de probabilité jointe des variables-partitions soit la loi de probabilité jointe empirique, observée dans les vrais co-clusters. Plus les données sont nombreuses, plus la loi empirique est proche de la loi théorique ;
- *l'information mutuelle normalisée entre les variables-partitions obtenues avec MODL (NMI MODL)* ;
- *l'information mutuelle normalisée entre les variables-partitions obtenues par maximisation de la modularité (NMI Modularité).*

Nous étudions ces mesures sur les graphes avec une proportions d'arêtes inter-cliques de 80%.

Peu importe le nombre d'arêtes engendrées, la NMI théorique reste constante puisque elle est indépendante du volume de données engendrées et que les probabilités de connexion des clusters sont inchangées pour chaque graphe engendré. La NMI empirique, quant à elle, décroît et finit par converger vers la NMI théorique à partir de 2 000 arêtes engendrés. Cela signifie que pour un nombre inférieure à 2 000 arêtes, la répartition des arcs observée dans les biclusters est différente et plus contrastée que la répartition théorique engendrée. Ainsi, en ayant tiré moins de 2 000 arêtes, les données engendrées ne sont pas en nombre suffisant pour faire apparaître la structure utilisée pour construire le graphe. La NMI de la structure de biclustering obtenue par MODL est nulle jusqu'à 2 000 arêtes engendrés, après quoi elle converge vers la NMI

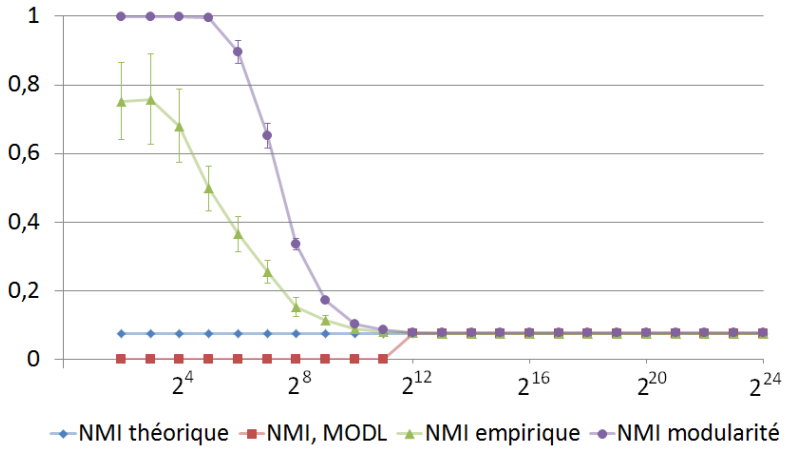


FIGURE 3.6 – Informations mutuelles normalisées moyennes théoriques, empiriques, obtenues avec MODL et par maximisation de la modularité, en fonction du nombre d'arcs engendrés. Les graphes sont composés de 100 nœuds et 10 clusters. La proportion d'arêtes dans les cliques est $p = 80\%$.

théorique. Avec moins de 2 000 arcs, aucune structure n'est construite. Après, la vraie structure est retrouvée : le régime asymptotique est atteint. Enfin, en ce qui concerne la NMI entre les variables-partitions obtenues par maximisation de modularité, elle est supérieure à la NMI empirique pour moins de 2 000 arêtes engendrées et converge vers la NMI théorique après. Cela signifie que la partition du graphe obtenue par maximisation de la modularité est meilleure que la partition théorique appliquée sur les données engendrées.

Au-delà de 2 000 arêtes engendrées, la maximisation de modularité et MODL retrouvent les bons clusters. Boullé (2012, EGC) a démontré la convergence asymptotique du critère MODL vers l'information mutuelle entre les variables. Cette propriété de l'approche est ici vérifiée empiriquement. Pour moins d'arêtes engendrés les deux approches produisent des résultats différents car elles n'ont pas les mêmes objectifs : MODL cherche à inférer les paramètres du modèle génératif des données alors que la maximisation de modularité cherche la meilleure partition empirique des données observées. Les performances des deux approches sont donc difficilement comparables.

3.2.5 Expérimentations sur des graphes simples orientés

Dans cette partie, nous construisons des graphes artificiels simples orientés. Les graphes sont engendrés suivant un modèle génératif infini, similaire à celui défini par Kemp et Tenenbaum (2006). Nous comparons l'approche MODL et une approche de blockmodeling stochastique sur les graphes construits.

Construction des données Le nombre de nœuds est égal à 100, et leur répartition dans les clusters suit un processus du restaurant chinois (prc) de

paramètre de concentration γ. Ainsi il est inutile de fixer le nombre et l'effectif des clusters. Dans l'approche de Kemp et Tenenbaum (2006), la matrice de connexion des clusters est tirée suivant une loi Beta. Construire des graphes avec cette répartition d'arcs dans les co-clusters produirait des graphes très différents avec un même paramétrage du modèle. De plus, rien ne garantit que des co-clusters contigus aient des probabilités de connexion différentes. De ce fait, nous préférons engendrer des graphes avec une structure simple à analyser. La matrice de probabilité de connexion entre les nœuds Θ est construite de manière à ce que les arcs dans les blocs de la partie triangulaire inférieure de la matrice d'adjacence aient une probabilité p d'être engendrés, que les arcs de la diagonale aient une probabilité $\frac{p}{2}$ et les arcs de la partie triangulaire supérieure aient une probabilité nulle. Enfin, pour chaque couple de nœuds (x_{i_1}, x_{i_2}), un arc $e_{i_1 i_2}$ est tiré suivant un tirage de Bernoulli de paramètre $\theta_{i_1 i_2}$, la probabilité de connexion des nœuds x_{i_1} et x_{i_2}. Les données sont engendrées suivant le protocole suivant :

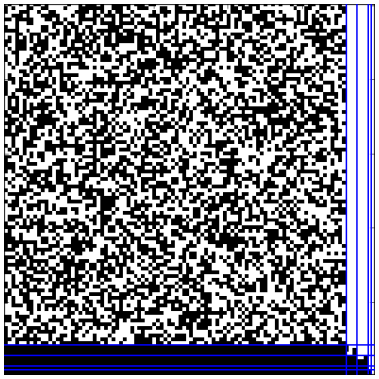
- choix du nombre de nœuds : $n = 100$;
- répartition des n nœuds dans les clusters : $X^M \sim prc(n, \gamma)$;
- construction de la matrice de probabilités de connexion des clusters :

$$\Theta = \begin{cases} \theta_{i_1 i_2} = \dfrac{p}{2} & \text{si } x_{i_1}, x_{i_2} \text{ sont dans le même cluster.} \\ \theta_{i_1 i_2} = p & \text{si } x_{i_1}, x_{i_2} \text{ sont dans des clusters différents et } i_1 > i_2 \\ \theta_{i_1 i_2} = 0 & \text{sinon.} \end{cases}$$

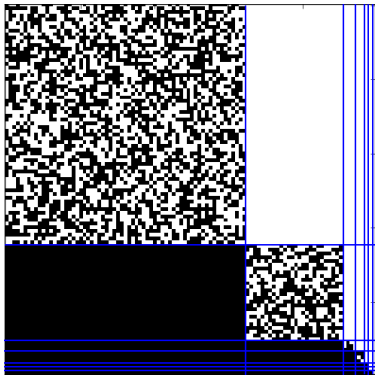
- entre tout couple de nœuds (x_{i_1}, x_{i_2}), un arc $e_{i_1 i_2}$ est tiré suivant une loi de Bernoulli :$\forall (x_{i_1}, x_{i_2}), e_{i_1 i_2} \sim Bernoulli(\theta_{i_1 i_2})$.

Les données sont engendrées pour différents paramètres γ et probabilités de connexion des clusters p. Plus le paramètre de concentration γ du processus du restaurant chinois est petit, moins les clusters sont nombreux et équilibrés. Au contraire, lorsque $\gamma \longrightarrow \infty$, on obtient un nœud par cluster. L'impact du paramètre de concentration du processus du restaurant chinois sur la répartition des nœuds dans les clusters est illustré par la figure 3.7. Nous prenons quatre valeurs différentes du paramètre γ dans les expérimentations : 10^{-2}, 10^{-1}, 1 et 10. Nous faisons varier la probabilité p de 0,05 à 1 par pas de 0,05. Pour chaque combinaison de p et γ, cent graphes différents sont engendrés.

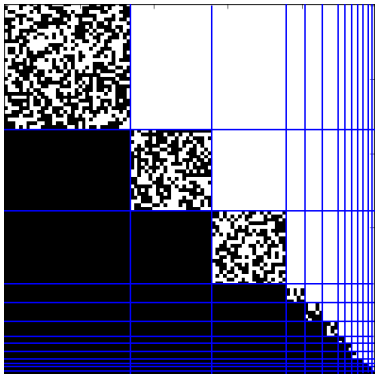
Expérimentations Nous comparons deux approches de partitionnement de graphe dans cette partie : le blockmodeling stochastique et l'approche MODL. La manière dont sont engendrées les données est favorable au blockmodeling stochastique car le modèle génératif des données est similaire au modèle inféré par cette méthode. Les deux approches cherchent des co-clusters d'arcs de



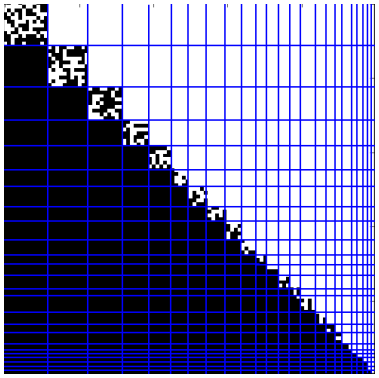
(a) Paramètre de concentration $\gamma = 10^{-2}$



(b) Paramètre de concentration $\gamma = 10^{-1}$



(c) Paramètre de concentration $\gamma = 1$



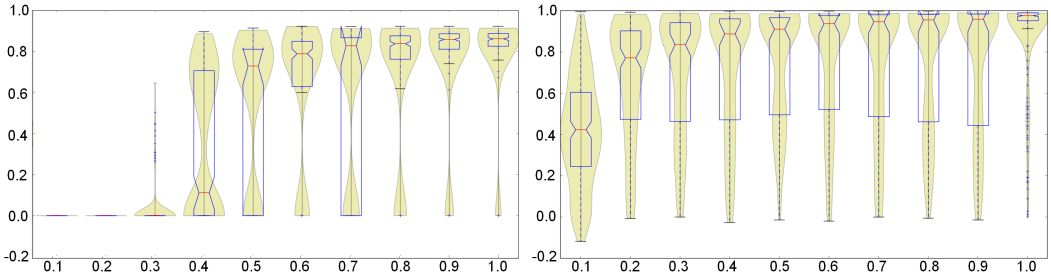
(d) Paramètre de concentration $\gamma = 10$

FIGURE 3.7 – Exemples de données engendrées suivant le modèle génératif avec différents paramètres de concentration du processus du restaurant chinois γ. La probabilité p vaut 1 dans les quatre cas.

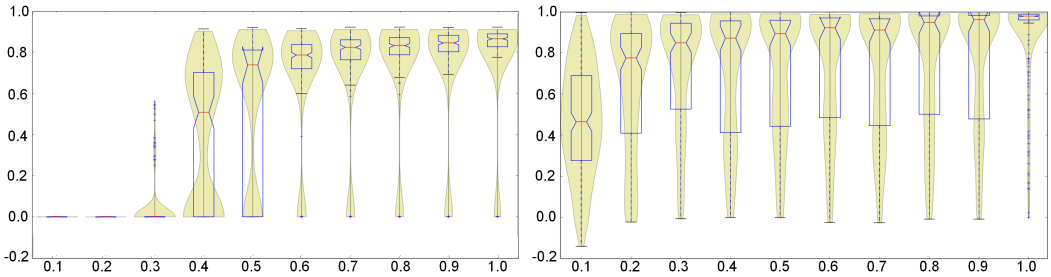
densités homogènes. L'hypothèse est faite que ces arcs sont engendrés suivant un modèle génératif. Les deux approches cherchent à en inférer les paramètres. L'approche MODL est implémentée dans le logiciel *Khiops*, qui a été utilisé pour cette étude. Pour le blockmodeling stochastique, nous utilisons l'approche de Latouche *et al.* (2010), implémentée dans le logiciel *Mixer* (Ambroise *et al.*, 2010). Dans cette implémentation, il est nécessaire de spécifier en paramètre les nombres de clusters minimal et maximal. De manière à pouvoir comparer les deux approches, nous spécifions un nombre de clusters pouvant aller de 1 à 100.

Résultats Dans les graphiques de la figure 3.8, on utilise de nouveau les diagrammes en violon pour montrer l'évolution de l'indice de Rand ajusté entre la vraie partition du graphe et les partitions obtenues par les approches comparées (MODL et le blockmodeling stochastique) en fonction de la probabilité de

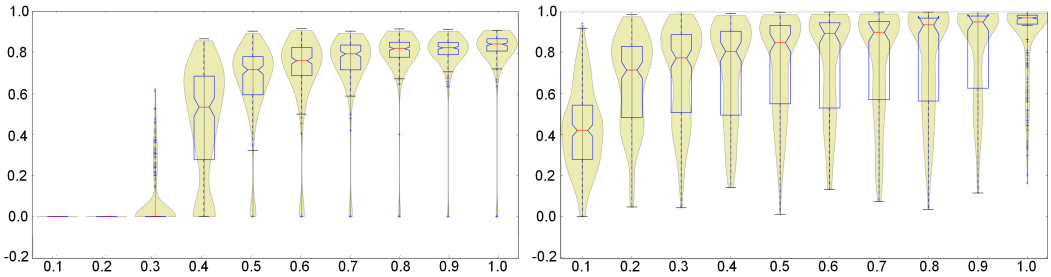
connexion des clusters. On ne s'intéresse pas au nombre de clusters car il peu différer d'un graphe à l'autre pour une même combinaison de p et γ.



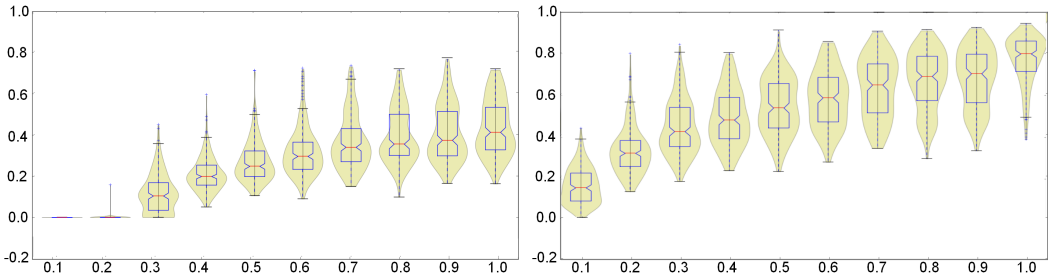
(a) $\gamma = 10^{-2}$; MODL (b) $\gamma = 10^{-2}$; SBM



(c) $\gamma = 10^{-1}$; MODL (d) $\gamma = 10^{-1}$; SBM



(e) $\gamma = 1$; MODL (f) $\gamma = 1$; SBM



(g) $\gamma = 10$; MODL (h) $\gamma = 10$; SBM

FIGURE 3.8 – Indice de Rand ajusté en fonction de p la probabilité de connexion des clusters. L'expérience compare les résultats trouvés par MODL et par une approche de blockmodeling stochastique (SBM) pour quatre différents paramètres de concentration γ du processus du restaurant chinois.

Pour les valeurs 10^{-2}, 10^{-1} et 1 du paramètre de concentration du processus du restaurant chinois, l'indice de Rand ajusté évolue de manière similaire pour chacune des approches. Lorsque la probabilité de connexion des clusters est faible, l'approche MODL ne produit aucune partition du graphe. À partir de $p = 0{,}25$, des structures sont retrouvées. L'ARI median croit avec la probabilité de connexion des nœuds jusqu'à une valeur autour de 0,9, qui n'est jamais excédée. Cela signifie que MODL trouve des partitions qui coïncident de mieux en mieux avec la partition du graphe engendrée. En faisant croître la probabilité de connexion de nœuds, on fait également croître le nombre d'arcs, ce qui explique ce gain d'ARI. L'approche de blockmodeling stochastique nécessite beaucoup moins d'arcs pour inférer une structure. Dès $p = 0{,}05$, l'ARI est fort. L'approche infère donc une partition pertinente vis-à-vis de la partition engendrée, même avec très peu de données. L'ARI médian obtenu avec l'approche de blockmodeling stochastique est meilleur que l'ARI obtenu avec MODL, mais la variance des résultats est très forte. On obtient des ARI négatifs pour certains graphes, ce qui signifie que la partition obtenue est pire qu'une partition aléatoire. Pour des fortes valeurs de p, les partitions inférées par le blockmodeling stochastique sont de qualité supérieure aux partitions trouvées par l'approche MODL. Le blockmodeling stochastique parvient à trouver la bonne partition (ARI de un), ce que MODL ne parvient pas à faire. On note malgré tout que les partitions de MODL ont tendance à être plus parcimonieuses en nombre de clusters que les partitions obtenues par application du blockmodeling stochastique.

Pour une valeur $\gamma = 10$ du paramètre de concentration du processus du restaurant chinois, l'évolution de l'indice de Rand ajusté médian est quasi linéaire pour les deux approches. Le blockmodeling stochastique parvient beaucoup mieux à reconnaître la structure que MODL. La figure 3.7.(d) montre à quel point la structure est complexe : il y a un grand nombre de clusters très peu peuplés. Pour retrouver ce type de structures, l'approche MODL a besoin d'un grand nombre de données, ce qui n'est pas le cas ici car les arcs sont simples, ce qui implique que leur nombre est limité par le nombre de nœuds. Pour les graphes construits avec ces paramètres, le blockmodeling stochastique produit systématiquement une partition alors que MODL, dans la plupart des cas, ne parvient pas à produire des clusters.

Pour l'ensemble des diagrammes de la figure 3.8, on remarque que la variance de l'ARI entre la partitions trouvées par le blockmodeling stochastique et les partitions engendrées est bien plus forte que la variance de l'ARI entre les partitions trouvées par MODL et les partitions engendrées. L'approche MODL est bien plus parcimonieuse que le blockmodeling stochastique. MODL a donc tendance à sous-partitionner le graphe mais avec une qualité constante. Au contraire l'approche de blockmodeling stochastique trouve la plupart du temps beaucoup plus de clusters. Dans certains cas, l'ARI entre la partition trouvée et la partition engendrée vaut 1, ce qui signifie que la partition inférée est la partition engendrée. Mais dans d'autres cas, on a un ARI négatif : la partition produite est pire qu'une partition aléatoire. L'approche MODL est donc plus

« prudente » que le blockmodeling stochastique dans le sens où elle ne fait pas émerger de structures non pertinentes, mais nécessite un plus grand nombre d'observations pour trouver des clusters.

Pour $p = 1$ et pour les deux approches comparées, l'ARI ne vaut jamais 1. Cela signifie que la structure trouvée n'est pas celle qui a été engendrée. Les clusters ne contenant qu'un seul nœud sont rarement trouvés. C'est ce qui explique cette valeur de l'ARI. Boullé (2012, EGC) a montré que dans le cas d'un biclustering de variables nominales, la répartition des arcs dans les biclusters trouvée par l'approche MODL converge asymptotiquement vers la vraie loi de probabilité jointe des variables. En augmentant le nombre d'arcs de manière à construire un multigraphe et en conservant la répartitions des arcs, on inférerait la vraie partition du graphe. Dans l'étude comparative sur des multigraphes de la section 3.3.3, on a identifié trois phases dans l'évolution de l'ARI en fonction du nombre d'arcs engendrés : une phase initiale où aucune structure n'est trouvée, une phase transitoire où une structure partielle est inférée et le régime asymptotique où la vraie structure est obtenue. Dans les expérimentations présentées ici, on retrouve les deux premières phases. Au début, aucune structure n'émerge car les observations sont trop peu nombreuses, puis à partir de $p = 0{,}25$ (degré moyen des nœuds de 12) la structure commence à apparaître. Le nombre d'arcs engendrés n'est cependant jamais suffisant pour trouver intégralement la structure du graphe.

De manière générale, l'approche de blockmodeling stochastique produit des meilleurs résultats que MODL sur ces données. Cependant, les données ont été engendrées de manière à être favorable à l'approche de blockmodeling stochastique. L'utilisation de cette approche est surtout bénéfique lorsque les probabilités de connexions des clusters sont faibles. Dans le cas contraire, on préfère utiliser MODL car les résultats sont plus stables. En termes d'algorithmique, l'approche MODL est plus rapide que le blockmodeling stochastique : le temps de calcul pour réaliser l'ensemble de ces expériences est de quelques minutes pour MODL et d'environ deux jours pour le blockmodeling stochastique.

Les expérimentations des sections 3.2.4 et 3.2.5 ont montré que l'approche MODL est adaptée au partitionnement de graphes. Dans MODL, l'unité statistique est l'arc. L'application de cette approche est donc plus pertinente pour traiter des multigraphes que des graphes simples. Les données ont été engendrées artificiellement suivant deux protocoles menant à des structures très différentes. Dans les deux cas, l'approche MODL s'est montrée robuste et performante, ce qui montre la flexibilité de son utilisation dans les problèmes de clustering de nœuds dans les graphes. Nous n'avons pas étudier le problème des clusters chevauchants dans cette section, l'approche MODL n'est actuellement pas conçu pour détecter ce type de motifs. Néanmoins, l'introduction d'une nouvelle formalisation du problème pourrait permettre de détecter ce types de structures.

3.3 Le triclustering hétérogène pour l'analyse de graphes temporels

Dans la continuité de l'analyse de graphes, on propose maintenant de considérer une évolution temporelle des clusters. Le co-clustering permet de réaliser une partition simultanée de plusieurs variables descriptives des données. Dans le cas des graphes simples, les deux variables sont les nœuds source et les nœuds cible. Dans le cas de graphes temporels, il suffit d'ajouter une troisième variable qui est le temps. On parlera alors de co-clustering en trois dimensions ou encore de *triclustering*. Dans cette partie, nous formalisons le problème de la segmentation de graphes temporels pour le traiter par le co-clustering et menons des expérimentations sur des données artificielles. Des expérimentations sur des données réelles du réseau de vélo en libre service de Londres ont été réalisées dans Guigourès *et al.* (2012).

3.3.1 État de l'art des approches de clustering de graphes temporels

Contrairement à l'étude des graphes statiques, l'analyse des *graphes temporels* est relativement récente. Un graphe temporel est un graphe qui évolue sur une période de temps définie pouvant modéliser, par exemple, des échanges d'emails sur une année ou encore des transactions commerciales pendant deux mois. Plusieurs types d'études peuvent être menées sur des données de cette nature. Leskovec *et al.* (2005) s'intéressent à l'évolution des paramètres caractéristiques du graphe au cours du temps, comme le nombre d'arcs, le degré moyen des nœuds ou encore le diamètre du graphe, c'est-à-dire la distance maximale entre deux nœuds du graphe. Xing *et al.* (2010) utilisent l'approche de blockmodeling stochastique d'Airoldi *et al.* (2008) pour étudier l'appartenance des nœuds aux différents clusters et étudient l'évolution de cette appartenance au cours du temps. Ici, nous cherchons à déterminer la meilleure partition du graphe temporel, c'est-à-dire le meilleur clustering de nœuds et le meilleur découpage temporel permettant de mettre en évidence des ruptures dans l'évolution du graphe.

Les premières études des graphes temporels ont cherché à adapter les critères déterministes utilisés pour déterminer la structure des graphes statiques. Hopcroft *et al.* (2004) se sont les premiers intéressés à l'évolution des clusters de nœuds dans le temps. Dans leur approche, le graphe temporel est étudié comme une séquence de graphes statiques. Pour chaque graphe statique, une classification hiérarchique ascendante est utilisée afin de retrouver la structure de clustering. La mesure de similarité est une mesure de cosinus qui est couramment employée pour la fouille de données textuelles (Li et Jain, 1998). Une fois, les clusters obtenus, des nœuds sont aléatoirement retirés et les clusters peu affectés par cette perturbation du graphe sont appelés *Communautés naturelles*. L'évolution de ces clusters est étudiée d'un graphe statique à l'autre de la

séquence mettant en évidence la naissance, le peuplement et la mort du cluster.

Palla *et al.* (2007) proposent une adaptation de la méthode de *clique percolation* (Palla *et al.*, 2005) pour les graphes temporels. Le principe est également d'étudier une séquence de graphes statiques. Le graphe au temps t est fusionné avec le graphe au temps $t+1$, et la méthode de clique percolation (capable de détecter des clusters chevauchants qui n'apparaissent que rarement) est appliquée sur le graphe obtenu. Le cluster à $t+1$ qui a le chevauchement maximal avec un cluster au temps t est considéré comme la version évoluée du cluster au temps t. Ainsi, on peut suivre l'évolution des clusters au cours du temps et quantifier leur stabilité en utilisant un indice de Jaccard (Duda *et al.*, 2001) entre les clusters au temps t et les clusters au temps $t+1$.

Sun *et al.* (2007) ont utilisé une méthode basée sur la théorie de l'information nommée *GraphScope*. Cette méthode est assez proche de MODL dans le sens où elle minimise un critère basée sur une approche MDL (Grünwald, 2007), avec un logarithme de probabilité a priori codant les paramètres de modélisation et un logarithme de la vraisemblance du modèle codant l'information présente dans la partition du graphe. Cependant, *GraphScope* calcule le clustering des nœuds et la discrétisation temporelle en deux étapes indépendantes, contrairement à MODL qui engendre le graphe image en une seule étape.

Les méthodes précédemment citées calculent la discrétisation temporelle et les clusters dans deux phases distinctes. Par conséquent, les corrélations entre la structure de clustering des nœuds et le temps ne sont pas capturées. De plus, considérer un graphe temporel comme une succession de graphes statiques requiert une pré-discrétisation du temps, ce qui pourrait faire disparaître des motifs très fins des données. Enfin, comme le souligne Fortunato (2010), ces méthodes sur des données bruitées trouvent des clusterings de nœuds très différents d'un graphe de la séquence à l'autre, et donc l'évolution temporelle de la structure sous-jacente du graphe n'est pas retrouvée.

En réalisant une partition simultanée de trois variables, le triclustering est une bonne solution pour intégrer la discrétisation temporelle et la construction des clusters dans une même phase. Zhao et Zaki (2005) ont déjà utilisé le triclustering pour étudier l'évolution de données d'expression de gènes. Cependant leur approche est basée sur la détection de motifs locaux et n'est donc pas comparable à MODL qui est, quant à elle, mieux adaptée à l'analyse de la structure globale et donc des graphes temporels.

3.3.2 MODL pour les graphes temporels

Le graphe est considéré comme un ensemble d'arcs décrits par trois variables : le nœud source, le nœud cible et une estampille temporelle. Nous considérons l'ensemble des nœuds constant dans le temps et l'étude se fait sur une période déterminée. Le temps est donc une variable temporelle définie sur un intervalle $[T_{min}, T_{max}]$. Notons que si on observe l'apparition d'un acteur dans le graphe, celui-ci est traité comme s'il était présent dès T_{min} mais inactif jusqu'à son apparition. De la même manière s'il disparaît du graphe, le nœud reste présent

dans le graphe mais devient inactif. Avoir un ensemble des nœuds constants n'est pas un problème puisque les observations sont les arcs dans l'approche MODL.

Un graphe temporel est donc un graphe dont l'ensemble des arcs évolue au cours du temps. Le temps est une variable continue qui est discrétisée en intervalles. Le graphe image est une séquence de graphes statiques, représentations synthétiques du graphe initial projeté sur chaque intervalle de temps. Le tableau 3.2 donne les notations.

L'utilisation du triclustering pour ce type d'études ne permet pas de faire apparaître une évolution du contenu des clusters. Ainsi, si certains nœuds passent d'un cluster 1 à un cluster 2 au cours de la période étudiée, on considère qu'il y a trois clusters : les nœuds changeant de clusters sont groupés dans un cluster 3, similaire au cluster 1 avant la migration des nœuds et similaire au cluster 2 après la migration des nœuds.

Le graphe $\mathcal{D}$	**Le graphe image $\mathcal{M}$**
X_1 les nœuds source n_1 le nombre de nœuds source $m_{i_1..}$ degré du nœud source i_1	X_1^M la partition des nœuds source k_1 le nombre de clusters source $m^C_{i_1..}$ degré du cluster source i_1
X_2 les nœuds cible n_2 le nombre de nœuds cible $m_{.i_2.}$ degré du nœud cible i_2	X_2^M la partition des nœuds cible k_2 le nombre de clusters cible $m^C_{.i_2.}$ degré du cluster cible i_2
X_3 le temps	X_3^M la discrétisation du temps k_3 le nombre d'intervalles de temps $m^C_{..i_3}$ nombre d'arcs observés dans l'intervalle i_3
m le nombre d'arcs	$k = k_1k_2k_3$ le nombre de triclusters $m^C_{i_1i_2i_3}$ effectif du tricluster formé par i_1, i_2 et i_3

TABLE 3.2 – Notations pour les graphes

Les paramètres du graphe image sont les mêmes pour un graphe temporel que pour un graphe statique, étendus avec les paramètres de la discrétisation temporelle :

1. **les paramètres de la partition pour un graphe statique**. Voir section 3.2. Notons que, dans le cas de graphes temporels, la répartition des arcs est spécifiée dans les triclusters formés par la partitions des nœuds source, cible et la discrétisation du temps ;
2. **le nombre d'intervalles de temps**. La variable temporelle est continue, on a donc un nombre potentiellement infini d'estampilles temporelles observables. Cependant, afin de pouvoir fixer un a priori sur le nombre d'intervalles de temps, on fait l'hypothèse que le nombre d'intervalles est compris entre 1 et m le nombre d'arcs du graphe. On obtient un

seul intervalle dans le cas d'un graphe stationnaire, c'est-à-dire avec une structure sous-jacente qui n'évolue pas au cours du temps.

Ces paramètres spécifiés, le critère décrit en section 2.2.5 peut se simplifier comme décrit dans l'équation 3.3.

Définition 18. *Le meilleur modèle $\mathcal{M}$ (ou graphe image dans le cas des graphes) est obtenu en maximisant le critère MODL ξ.*

$$
\begin{aligned}
\xi(\mathcal{M}) = & \log n_1 + \log n_2 + \boldsymbol{\log m} \\
& + \log B(n_1, k_1) + \log B(n_2, k_2) \\
& + \log \binom{m+k-1}{k-1} \\
& + \sum_{i_1=1}^{k_1} \log \binom{m^C_{i_1..} + n^C_{i_1..} - 1}{n^C_{i_1..} - 1} + \sum_{i_2=1}^{k_2} \log \binom{m^C_{.i_2.} + n^C_{.i_2.} - 1}{n^C_{.i_2.} - 1} \\
& + \log m! - \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} \sum_{\boldsymbol{i_3=1}}^{\boldsymbol{k_3}} \log m^C_{i_1 i_2 \boldsymbol{i_3}}! \\
& + \sum_{i_1=1}^{k_1} \log m^C_{i_1..}! - \sum_{i_1=1}^{n_1} \log m_{i_1..}! + \sum_{i_2=1}^{k_2} \log m^C_{.i_2.}! - \sum_{i_2=1}^{n_2} \log m_{.i_2.}! \\
& + \sum_{\boldsymbol{i_3=1}}^{\boldsymbol{k_3}} \boldsymbol{\log m^C_{..i_3}!}
\end{aligned}
\tag{3.3}
$$

Les éléments en gras sont les éléments du critère adapté aux graphes temporels, ajoutés afin de pouvoir trouver la structure temporelle des graphes qui évoluent au cours du temps. Notons que peu d'éléments sont ajoutés : un terme de probabilité a priori à la première ligne permettant d'affecter une probabilité a priori sur le nombre d'intervalles, ainsi qu'un terme de vraisemblance en dernière ligne décrivant la probabilité d'observer les données pour une discrétisation donnée.

3.3.3 Expérimentations sur des données artificielles

On a vu précédemment que MODL retrouvait des structures de biclustering diversifiées à l'aide d'expérimentations sur des données produites suivant différents modèles génératifs. Ici, on s'intéresse à la dimension numérique du triclustering, c'est-à-dire l'évolution temporelle du graphe. On décide donc de fixer une partition des nœuds de manière relativement simple : 50 nœuds distribués de manière équilibrée dans 5 clusters.

Comme pour les graphes modulaires engendrés dans la section 3.2.4, on a des proportions p d'arcs dans les clusters et $1-p$ entre les clusters. Ces proportions évoluent au cours du temps. À T_{min}, la proportion d'arcs entre les clusters vaut 0,9 et la proportion d'arcs au sein des clusters vaut 0,1. Ces proportions sont

inversées à T_{Max}. Dans la terminologie des graphes, on démarre d'une structure de quasi-cocliques pour terminer sur une structure de quasi-cliques. Entre T_{min} et T_{max}, on propose une évolution linéaire de la structure, c'est-à-dire que la valeur des probabilités de connexion des nœuds évolue linéairement avec le temps. Comme dans la section 3.2.4, nous appliquons notre approche de triclustering pour un nombre croissant d'arcs engendrés. Pour simplifier l'interprétation, on prend $T_{min} = 0$ et $T_{max} = 1$. Les estampilles temporelles x_{i_3} prennent donc des valeurs sur $[0, 1]$. Dans un second temps, nous détruisons la structure temporelle du graphe. La loi de la variable *temps* est cette fois uniforme, ce qui en fait un graphe stationnaire

Évolution linéaire des graphes Étudions un graphe temporel dont les probabilités de connexion entre les nœuds évoluent linéairement au cours du temps. Le processus suivant est utilisé pour engendrer les données :

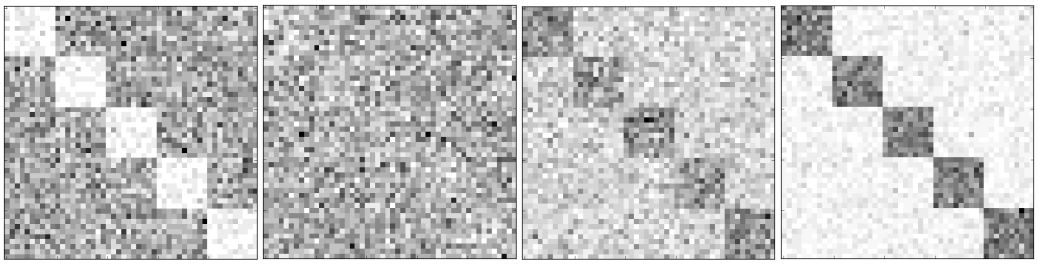
- choix du nombre de nœuds : $n = 50$;
- choix du nombre de clusters : $k = 5$;
- répartition des n nœuds dans les k clusters :
  $X^M = \{c_i, \forall i = 1..k, |c_i| = 10\}$;
- construction de la matrice de probabilités de connexion des nœuds :

$$\Theta(x_{i_3}) = \begin{cases} \theta_{i_1 i_2} = \dfrac{0{,}9x_{i_3} + 0{,}1(1 - x_{i_3})}{k} & \text{si } x_{i_1}, x_{i_2} \text{ dans le même cluster.} \\ \theta_{i_1 i_2} = \dfrac{0{,}1x_{i_3} + 0{,}9(1 - x_{i_3})}{k(k-1)} & \text{sinon ;} \end{cases}$$

- tirage aléatoire des estampilles temporelles entre 0 et 1 : $X_3 \sim \mathcal{U}(0, 1)$;
- tirage des arcs : $\forall x_{i_3} \in X_3, e_{i_1 i_2 i_3} \sim multinomiale\left(1, \dfrac{\Theta(x_{i_3})}{\sum_{i_1,i_2} \theta_{i_1 i_2}(x_{i_3})}\right)$.

La figure 3.9 permet de mieux visualiser l'évolution des motifs. Vers $x_{i_3} = 0$, les arcs sont denses en dehors de la diagonale, c'est la structure de quasi-cocliques. Aux alentours de $x_{i_3} = 0, 2$, la proportion d'arêtes est la même dans tous les co-clusters, on a donc un graphe aléatoire. Enfin, plus on s'approche de $x_{i_3} = 1$, plus on voit les arcs se densifier dans les biclusters diagonaux : on converge vers la structure en quasi-cliques.

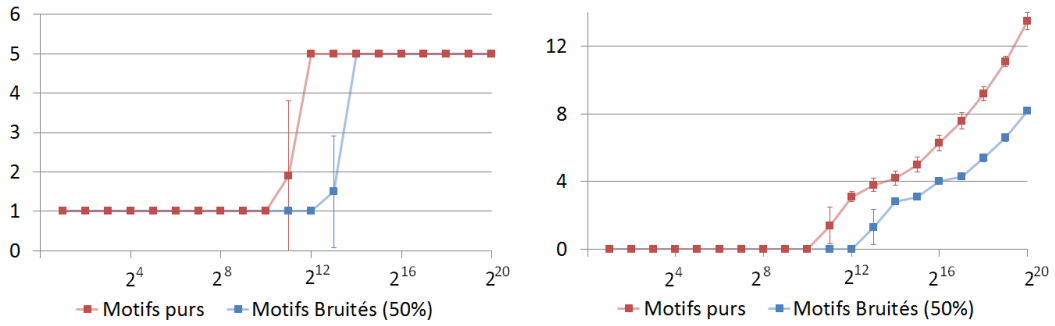
Les arcs sont engendrés par puissances de deux de 1 à environ 10^6. Pour un nombre d'arcs donné, vingt graphes sont construits : dix avec des motifs purs et dix avec 50% de bruit. L'ajout de bruit dans les données consiste à réattribuer aléatoirement un nœud source, un nœud cible et une estampille temporelle à 50% des arcs choisis au hasard. Les figures 3.10a et 3.10b présentent le nombre



(a) $[0, \epsilon]$ (b) $[0.2, 0.2 + \epsilon]$ (c) $[0.5, 0.5 + \epsilon]$ (d) $[1 - \epsilon, 1]$

FIGURE 3.9 – Graphe temporel projeté sur 4 intervalles de temps. $\epsilon = 10^{-2}$

de clusters moyen et le nombre d'intervalles moyen retrouvés en appliquant l'approche MODL.



(a) Nombre de clusters en fonction du nombre d'arcs engendrés

(b) Nombre d'intervalles en fonction du nombre d'arcs engendrés

FIGURE 3.10 – Évolution des nombres moyens de clusters de nœuds et d'intervalles de temps en fonction du nombre d'arcs engendrés dans des graphes avec une évolution linéaire.

Pour un faible nombre d'arcs engendrés (en dessous 2 000), la méthode ne parvient pas à retrouver la structure sous-jacente du graphe. Comme il a été vu dans les expérimentations de la section 3.2.4, lorsque le nombre d'arcs est trop faible pour que la structure engendrée puisse être retrouvée, MODL se montre robuste en ne créant aucune partition du graphe.

Entre 2 000 et 4 000 arcs, l'approche commence à détecter les clusters et à segmenter la variable temporelle. Cependant le nombre de clusters obtenus est inférieur au nombre de clusters engendrés. Certains clusters n'ont pu être discernés : il s'agit du régime transitoire avant le régime asymptotique.

Enfin, au-delà de 4 000 arcs, les clusters sont tous retrouvés et la méthode ne crée aucun cluster qui n'ait pas été engendré. Concernant la variable temporelle, la taille de sa partition croit régulièrement avec le nombre d'arcs. Au final, la structure évoluant de manière linéaire, plus il y a d'arcs engendrés, plus l'approche MODL discrétise finement la variable temporelle. Notons quand

même une régularité dans la discrétisation. Les intervalles sont de plus en plus fins au fur et à mesure qu'on s'éloigne de $x_{i_3} = 0{,}2$. En effet, plus les probabilités de connexion des biclusters sont contrastées, moins le nombre d'arcs nécessaire pour retrouver les motifs a besoin d'être haut. Ainsi, aux alentours de $x_{i_3} = 0{,}2$, le graphe est presque aléatoire et donc bien plus difficile à partitionner.

L'ajout de bruit dans le graphe temporel a rendu plus complexe l'analyse. MODL résiste bien au phénomène en ne créant pas de motifs incohérents. On remarque que le bruit retarde l'arrivée du régime asymptotique. L'approche a donc besoin de plus de données pour réussir à retrouver la structure sous-jacente du graphe temporel.

Graphes stationnaires Nous cherchons maintenant à montrer que l'approche MODL est en mesure de capturer la structure des données sur deux partitions et de ne pas segmenter les variables n'ayant aucune structure sous-jacente. Pour ce faire, nous proposons d'étudier des graphes stationnaires. Pour obtenir de tels graphes, les graphes précédents sont réutilisés. Les estampilles temporelles sont permutées d'arc en arc de manière aléatoire. Ainsi, la structure temporelle du graphe est totalement détruite mais la partition des nœuds demeure inchangée.

L'application de MODL sur les graphes ainsi engendrés ne fait apparaître aucune segmentation non-pertinente du temps en intervalles. De plus, elle permet de retrouver la partition des nœuds mais requiert un nombre plus important d'arcs par rapport à ce que nécessitait le même graphe avec une structure temporelle. En effet, en rendant le graphe stationnaire, on a mélangé la répartition d'arcs de chacune des valeurs de x_{i_3}, complexifiant ainsi beaucoup la structure de biclustering entre les nœuds.

3.4 Le triclustering hétérogène pour l'analyse de données fonctionnelles

L'analyse de graphes temporels est une application du triclustering de variables hétérogènes avec deux variables nominales et une variable numérique. Dans un tout autre domaine, on propose ici d'appliquer le triclustering au problème du clustering de données fonctionnelles. Dans ce cas, le triclustering est effectué sur des données décrites par deux variables continues et une variable nominale. Dans cette partie, nous rappelons la formalisation du problème du clustering de courbes pour le traiter par le co-clustering et menons des expérimentations sur des données artificielles. Des expérimentations sur des données réelles de consommation électrique sont présentées dans Boullé (2012, Pattern Recognition) et Boullé *et al.* (2013).

3.4.1 État de l'art des approches de clustering de données fonctionnelles

En analyse de données fonctionnelles (Ramsay et Silverman, 2005), les observations sont des fonctions (ou des courbes). Les données fonctionnelles sont présentes dans de nombreux domaines comme, par exemple, l'enregistrement des précipitations d'une station météorologique ou encore dans la surveillance de matériel, où chaque courbe est une série temporelle liée à une quantité physique enregistrée à fréquence spécifiée. Les méthodes d'analyse exploratoire pour les grandes bases de données fonctionnelles sont nécessaires dans de nombreuses applications pratiques comme par exemple, la surveillance de la consommation électrique (Hébrail *et al.*, 2010). Elles réduisent la complexité des données en combinant des techniques de clustering avec des méthodes d'approximation de fonction, modélisant, par exemple, un ensemble de données fonctionnelles par des *courbes prototypiques*, ensembles de segments linéaires ou de splines. Dans ce type d'approches, à la fois le nombre de prototypes et le nombre de segments sont des paramètres utilisateur.

Chamroukhi *et al.* (2010) et Hébrail *et al.* (2010) proposent de définir les motifs fonctionnels comme de simples fonctions telles que des fonctions indicatrices d'intervalles ou des polynômes simples : une courbe est approchée par une combinaison linéaire de ces fonctions simples.

Des approches Bayésiennes, comme celle proposée par Nguyen et Gelfand (2011), considèrent que l'ensemble des courbes peut être représenté par des courbes moyennes engendrées suivant un processus Gaussien et organisées en clusters. Alors que les modèles paramétriques utilisant un nombre fixe et fini de paramètres peuvent souffrir de sur/sous-apprentissage, des approches Bayésiennes non-paramétriques ont été proposées pour éviter ce problème. En utilisant un modèle de complexité non bornée, le sous-apprentissage est atténué, alors que l'approche Bayésienne de calcul ou d'approximation de la probabilité a posteriori des paramètres réduit le risque de sur-apprentissage (Teh, 2010). Au final, la loi des paramètres de clustering est obtenue en échantillonnant la probabilité a posteriori des paramètres en utilisant des méthodes d'inférence Bayésienne comme les *MCMC* (Neal, 2000) ou l'inférence variationnelle (Blei et Jordan, 2005). Suit un post-traitement permettant de choisir un clustering. Un a priori basé sur une loi de Dirichlet nécessite deux paramètres utilisateur : le paramètre de concentration et la loi de base. Pour un paramètre de concentration γ et un jeu de données contenant n courbes, l'espérance du nombre de clusters $\bar{k}$ est $\bar{k} = \gamma \log(n)$ (Wallach *et al.*, 2010). De ce fait, le paramètre de concentration a un impact significatif sur le nombre de clusters obtenus. Selon Vogt *et al.* (2010), il n'est pas possible d'estimer de manière fiable ce paramètre.

L'approche MODL est comparable aux approches basées sur les *processus de Dirichlet* dans le sens où elles estiment une probabilité a posteriori basée sur la vraisemblance et la probabilité a priori des paramètres d'un modèle. Cependant, MODL est intrinsèquement différent des méthodes basées sur les processus de Dirichlet. Ces approches sont Bayésiennes et estiment le

paramètre d'une distribution de clusterings, le clustering final étant obtenu par un post-traitement consistant, par exemple, à choisir le mode de la probabilité a posteriori ou encore en étudiant la matrice des co-occurrences. A contrario, MODL est une approche MAP, le modèle le plus probable est directement recherché en utilisant des algorithmes d'optimisation.

3.4.2 MODL pour les données fonctionnelles

Une courbe est un ensemble de points prenant des valeurs en abscisse sur X_1 et en ordonnée sur X_2. Chaque courbe est décrite par un identifiant qui fait office de troisième variable descriptive X_3. On a donc deux variables numériques X_1 et X_2 et une variable nominale X_3. L'unité statistique est le point : chaque point est décrit par une valeur en abscisse, en ordonnée et un identifiant de courbe. Le tableau 3.3 donne les notations.

Ensemble de courbes $\mathcal{D}$	**Modèle $\mathcal{M}$**
X_1 variable en abscisse	X_1^M la partition de l'axe des abscisses k_1 nombre d'intervalles en abscisse $m^C_{i_1 \cdot}$ nombre de points dans l'intervalle i_1
X_2 variable en ordonnée	X_2^M la partition de l'axe des ordonnées k_2 nombre d'intervalles en ordonnée $m^C_{\cdot i_2}$ nombre de points dans l'intervalle i_2
X_3 identifiant des courbes n_3 le nombre de courbes $m_{\cdot\cdot i_3}$ nombre de points de la courbe i_3	X_3^M la partition des courbes k_3 le nombre de clusters de courbes $m^C_{\cdot\cdot i_3}$ nombre de points du cluster de courbes i_3
m le nombre total de points	$k = k_1 k_2 k_3$ le nombre de triclusters $m^C_{i_1 i_2 i_3}$ effectif du tricluster formé par i_1, i_2 et i_3

TABLE 3.3 – Notations pour les graphes

L'utilisation du triclustering a ici pour but de découper l'axe des abscisses et des ordonnées en intervalles et de grouper les courbes dans des clusters. Ces segmentations caractérisent un modèle synthétique $\mathcal{M}$ inféré à partir des données $\mathcal{D}$. Les paramètres de ce modèle sont décrits ci-dessous :

1. **le nombre de clusters de courbes et d'intervalles**. De manière analogue aux graphes temporels, le nombre d'intervalles sur les axes des abscisses et des ordonnées, ainsi que le nombre de clusters de courbes sont a priori inconnus. Le nombre de clusters de courbes k_3 est uniformément distribué entre 1 et n_3, le nombre de courbes, et les nombres d'intervalles k_1 et k_2 entre 1 et m le nombre total de points ;

2. **la répartition des courbes dans les clusters**. Une fois le nombre de clusters k_3 connu, le nombre de courbes qui les peuplent est spécifié. Toutes les répartitions des n_3 courbes dans k_3 clusters, potentiellement vides, sont équiprobables ;

3. **la répartition des points dans les triclusters**. Avec les nombres d'intervalles k_1 et k_2 et le nombre de clusters de courbes k_3, on peut déduire le nombre de triclusters $k = k_1 k_2 k_3$. On considère que toutes les répartitions des m points sur les k triclusters sont équiprobables ;

4. **la répartition des points d'un cluster dans les courbes qu'il contient**. Les nombres de points dans les clusters de courbes sont déduits des effectifs des k triclusters. On considère équiprobables, pour chaque cluster, toutes les façons de distribuer les $m^C_{..i_3}$ points du cluster i_3 sur les $n^C_{..i3}$ courbes qu'il contient.

Ces paramètres étant spécifiés, le critère décrit en section 2.2.5 peut se simplifier pour être adapté au clustering de courbes, comme décrit dans l'équation 3.4.

Définition 19. *Le meilleur modèle $\mathcal{M}$ est obtenu en maximisant le critère MODL ξ.*

$$\begin{aligned}
\xi(\mathcal{M}) =& 2\log m + \log n_3 \\
&+ \log B(n_3, k_3) \\
&+ \log \binom{m+k-1}{k-1} \\
&+ \sum_{i_3=1}^{k_3} \log \binom{m^C_{..i_3} + n^C_{..i_3} - 1}{n^C_{..i_3} - 1} \\
&+ \log m! - \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\sum_{i_3=1}^{k_3} \log m^C_{i_1 i_2 i_3}! \\
&+ \sum_{i_3=1}^{k_3} \log m^C_{..i_3}! - \sum_{i_3=1}^{n_3} \log m_{..i_3}! \\
&+ \sum_{i_1=1}^{k_1} \log m^C_{i1..}! + \sum_{i_2=1}^{k_2} \log m^C_{.i_2.}!
\end{aligned} \tag{3.4}$$

Le critère de triclustering adapté aux courbes est très proches de celui des graphes temporels : les termes d'a priori et de vraisemblance sont les mêmes sauf qu'ils codent deux variables numériques et une variable nominale dans le cas des courbes, alors qu'ils codaient deux variables nominales et une numérique dans le cas des graphes temporels.

3.4.3 Expérimentations sur des données artificielles

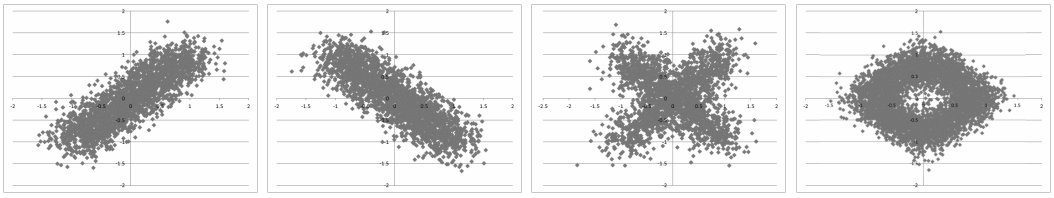
Comme dans les études précédentes, des données artificielles sont engendrées. Pour cela, plusieurs courbes sont construites suivant quatre motifs bruités, représentant les clusters de courbes. Les données sont engendrées suivant les fonctions suivantes :

- choix du nombre de courbes : $n_3 = 40$;
- choix du nombre de clusters : $k_3 = 4$;
- répartition des n_3 courbes dans les k_3 clusters :

  $X_3^M = \{c_{..i_3}, \forall i_3 = 1..k_3, |c_{..i_3}| = 10\}$;
- choix du niveau de bruit des courbes : $\varepsilon \sim \mathcal{N}(0, 0.25)$;
- définitions des motifs associés à chacun des clusters

$$F(x_{..i_3}, z) = \begin{cases} \text{Si } x_{..i_3} \in c_{..1} : & x_{i_1..} = z + \varepsilon \\ & x_{.i_2.} = z + \varepsilon \\ \text{Si } x_{..i_3} \in c_{..2} : & x_{i_1..} = z + \varepsilon \\ & x_{.i_2.} = -z + \varepsilon \\ \text{Si } x_{..i_3} \in c_{..3} : & x_{i_1..} = z + \varepsilon \\ & x_{.i_2.} = \alpha z + \varepsilon \text{ avec } \alpha \sim \mathcal{U}(\{-1, 1\}) \\ \text{Si } x_{..i_3} \in c_{..4} : & x_{i_1..} = (0.75 + \varepsilon)\cos(\pi(1 + z)) \\ & x_{.i_2.} = (0.75 + \varepsilon)\sin(\pi(1 + z)) \end{cases}$$

- tirage des points : $x_{..i_3} \sim \mathcal{U}(X_3)$, $z \sim \mathcal{U}(-1, 1)$, $x_{i_1..}, x_{.i_2.} = F(x_{..i_3}, z)$.



(a) $c_{..1}$ (b) $c_{..2}$ (c) $c_{..3}$ (d) $c_{..4}$

FIGURE 3.11 – Motifs utilisés pour engendrer les données

L'ensemble $\mathcal{D}$ de 10^5 points est également engendré. Chaque point est un triplet de valeurs $(x_{i_1..}, x_{.i_2.}, x_{..i_3})$ avec un identifiant de courbe choisi parmi les 40 valeurs de X_3, une valeur de X_1 et de X_2 engendrées suivant la loi correspondant à la courbe choisie.

Nous appliquons l'approche MODL adaptée au clustering de données fonctionnelles sur des sous-ensembles de données $\mathcal{D}$ de tailles croissantes.

L'expérience est renouvelée 10 fois pour chaque sous-ensemble de points ré-échantillonné à chaque fois. Le graphique de la figure 3.12 montre le nombre moyen de clusters de courbes et le nombre moyen d'intervalles pour un nombre de points m. Pour les petits sous-ensembles (en-dessous de 400 points), il n'y pas suffisamment de données pour découvrir des motifs significatifs, et la méthode produit un unique cluster contenant toutes les courbes et un seul intervalle pour X_1 et X_2 . À partir de 400 points, les nombres de clusters et d'intervalles commencent à croître. Finalement pour 25 points par courbe en moyenne, c'est-à-dire 1 000 points au total, la méthode retrouve les structures sous-jacentes et produits les quatre clusters de courbes.

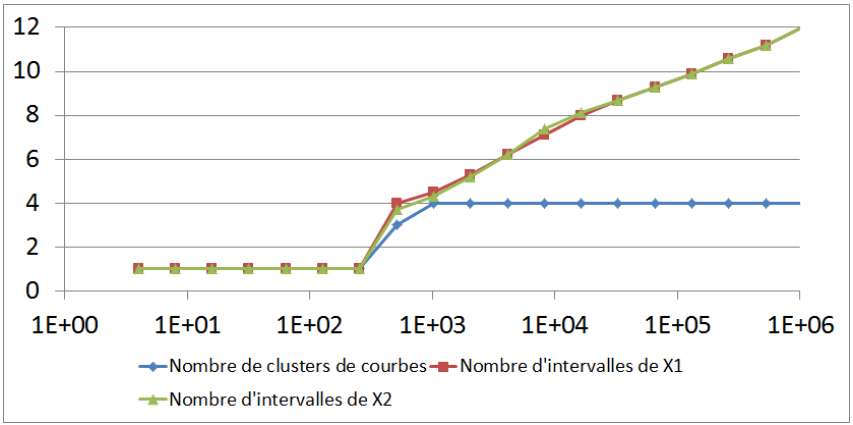


FIGURE 3.12 – Nombre de clusters de courbes, nombre d'intervalles de X_1 et nombre d'intervalles de X_2 en fonction du nombre de points m.

Bien que la méthode retrouve le véritable nombre de clusters, en-dessous de 2 000 points, les clusters peuvent ne pas être totalement purs et certaines courbes mal placées. Dans nos expériences, pour 1 000 points, en moyenne 2% des courbes sont mal placées, alors qu'avec 2 000 points, les courbes sont systématiquement classées dans le bon cluster. Notons qu'en augmentant la taille du sous-ensemble de points au-delà de 2 000 points, le nombre de clusters de courbes obtenu est constant et égal à quatre. A contrario, le nombre d'intervalles croit avec le nombre de points. Ceci montre le bon comportement asymptotique de la méthode : on retrouve le bon nombre de motifs et la méthode exploite la quantité croissante des données pour mieux approximer la forme des motifs.

Cette expérience permet de mettre en évidence une propriété intéressante : la méthode ne nécessite pas que les positions des valeurs des variables X_1 et X_2 soient les mêmes pour toutes les courbes. De plus, il est possible de détecter des clusters avec des multimodalités comme les clusters de courbes 3 et 4 par exemple. Dans ces exemples, pour un intervalle de X_1^M fixé, les observations (points des courbes) sont distribuées sur deux clusters de la variable-partition X_2^M. Dans ce genre de situations, la moyenne n'est pas représentative des courbes du cluster. MODL a donc un avantage sur les approches alternatives, qui cherchent à grouper des courbes les plus proches possibles de la courbe

moyenne, représentative du cluster. D'un point de vue pratique, la détection de multi-modalité est utile pour observer des régimes transitoires dans l'évolution des clusters de courbes. Une illustration de ce type de phénomènes est présentée dans Boullé *et al.* (2013) sur un enregistrement de consommation électrique.

3.5 Le d-clustering hétérogène pour l'analyse de données complexe

On a vu que l'utilisation du co-clustering pouvait être une réponse à des problèmes diversifiés tels que le partitionnement de graphes statiques, de graphes temporels ou encore le clustering de courbes. L'approche MODL ne se limite pas à du co-clustering en deux ou trois dimensions, mais peut également être appliquée à des données décrites par plusieurs dimensions. Dans cette section, nous verrons quelles peuvent être les applications de ce type d'extensions.

3.5.1 Intérêt et potentiel du d-clustering

On a vu précédemment que le co-clustering pouvait être un outil de recherche de la structure temporelle de l'évolution d'un objet comme un graphe. Lorsque l'information temporelle est riche sur les arcs, on peut souhaiter savoir quelle structure temporelle est la plus corrélée avec des changements de distributions d'arcs dans le graphe. Ainsi, lorsqu'on s'intéresse au trafic dans un réseau, routier par exemple, il peut être intéressant de savoir comment sont structurés les déplacements suivant l'heure de la journée, le jour de la semaine, ou encore la date dans l'année. On peut également se dire que la distribution du trafic dans les réseaux peut à la fois dépendre du jour de la semaine et de l'heure. C'est pourquoi, s'intéresser simultanément à plusieurs variables temporelles peut s'avérer informatif pour l'utilisateur. On a donc, dans ce type d'analyses, les nœuds source, les nœuds cible et autant de variables continues qu'on a de marqueurs temporels différents à étudier.

Un autre exemple est le clustering de courbes. On a vu que l'utilisation du co-clustering était adaptée aux problèmes de clustering de courbes dans des applications comme les analyses de données, météorologiques par exemple. Mais lorsque de telles études sont réalisées pour déterminer le climat d'une région à partir d'un ensemble de mesures prises dans plusieurs stations, on dispose de plusieurs informations, comme la température moyenne, les quantités de précipitations ou de chutes de neiges. On a donc une variable nominale identifiant les stations météorologiques, une variable temporelle, et autant de variables numériques qu'il y a de types de mesures qui ont été effectuées. L'utilisation du co-clustering en d dimensions permet ici de réaliser un clustering avec des variables descriptives variées.

3.5.2 MODL, un critère général pour le co-clustering

Le critère, dans sa définition générale de la section 2.2.5, est défini pour un co-clustering en d dimensions. L'approche est donc adaptée à l'étude de données complexes où les observations sont décrites par d variables. On peut donc traiter des données en exploitant toutes les variables, d'autant que si une variable n'apporte pas d'information, celle-ci n'est pas segmentée par l'approche, comme on a pu le voir dans le cas des graphes stationnaires.

Cependant, un nombre important de variables rend les données plus coûteuses à analyser : la probabilité a priori spécifiée dans le critère code la complexité du modèle de co-clustering, donc plus il y a de variables segmentées dans le modèle, plus le coût de codage est fort. Ainsi, si deux variables sont très corrélées, leur segmentation contribue pour beaucoup dans la vraisemblance du modèle. La segmentation des $d-2$ autres variables est coûteuse en terme d'a priori pour peu de vraisemblance gagnée. Si le nombre d'observations n'est pas suffisant, la segmentation risque d'être ignorée même si elle est informative. Il faut donc être prudent lorsqu'on effectue du co-clustering en d dimensions, en s'assurant que le nombre d'observations est suffisant pour exploiter au mieux l'information apportée par les variables, et surtout, limiter les variables trop corrélées en préférant en sélectionner une seule pour le co-clustering.

Ce phénomène a pu être observé dans l'analyse de comptes-rendus d'appels (Guigourès *et al.*, 2013). Il s'agit de la liste d'appels entre antennes à l'échelle d'un pays. On dispose de l'antenne émettrice, l'antenne réceptrice, la date et l'heure de l'appel. On peut donc espérer faire plusieurs types d'analyses sur ces données comme un partitionnement du réseau d'antennes (biclustering nominal) ou encore une segmentation du graphe temporels ou même un co-clustering en quatre dimensions en étudiant le réseau d'antennes, l'heure de l'appel et le jour de la semaine (quatre variables : antennes source, antennes cible, heure et jour de la semaine). Cependant, on observe que la corrélation entre l'antenne émettrice et l'antenne réceptrice est si forte qu'elle masque intégralement les informations de l'évolution temporelle du graphe. Dans ce cas, on préfère réduire la dimension en considérant que l'ensemble des antennes source et l'ensemble des antennes cible sont identiques. Dans ce cas là, on obtient une partition des antennes différente nous permettant d'exploiter la structure temporelle des données.

3.6 Bilan

Dans ce chapitre, nous avons proposé de modéliser trois problèmes à l'aide de l'approche MODL : le partitionnement de graphes, de graphes temporels et le clustering de courbes. Après avoir fait un rapide état de l'art, on a introduit une formalisation de chacun des problèmes, de manière à pouvoir les traiter avec une approche de co-clustering. Enfin des expérimentations sur des données artificielles ont montré la fiabilité de l'approche MODL et la pertinence

de son utilisation sur les problèmes étudiés. De plus, des expérimentations comparatives avec d'autres approches ont montré que l'approche MODL avait des performances comparables à celles de méthodes concurrentes sur des données artificielles favorables à ces dernières.

Dans le chapitre 4, nous allons proposer une méthodologie d'analyse des résultats et introduire différents outils d'analyse exploratoire des grilles de co-clustering. Dans le chapitre 5, ces utilisations du co-clustering sont appliquées sur une base de données issue de l'opérateur Orange.

4

Exploitation et exploration des résultats

On considère dans ce chapitre que le modèle de co-clustering est connu et optimal suivant le critère MODL. L'enjeu majeur de l'approche est d'extraire des données une information utile et interprétable pour l'utilisateur. Les résultats bruts, bien qu'optimaux selon le critère de l'approche MODL, ne sont pas nécessairement directement exploitables. C'est pourquoi plusieurs outils d'analyse exploratoire sont proposés. Une méthode de simplification du co-clustering avec une dégradation maîtrisée de la qualité du modèle est d'abord introduite afin de permettre une interprétation du co-clustering à plusieurs échelles ; ensuite, des indicateurs permettant l'étude des clusters et des individus les composant sont proposés afin de détecter les clusters intéressants et les valeurs représentatives de leur cluster.

Afin de simplifier l'introduction des différents outils d'analyse exploratoire, nous traitons dans un premier temps le cas du co-clustering à deux dimensions nominales. Nous proposons ensuite une généralisation à plus de deux dimensions, nominales et continues.

4.1 Définitions, notations et exemple illustratif

Nous introduisons dans cette section les notations qui sont utilisées par la suite. Dans le chapitre 2, on a vu que les données $\mathcal{D}$ peuvent être vues comme un ensemble de m points décrits par deux variables X_1 et X_2 prenant respectivement des valeurs dans les ensembles $\{v_{1.}, v_{2.}, ..., v_{n_1.}\}$ et $\{v_{.1}, v_{.2}, ..., v_{.n_2}\}$.

Ces valeurs sont regroupées en k_1 et k_2 groupes suivant les paramètres du modèle $\mathcal{M}$. Les variables-partitions issues du modèle $\mathcal{M}$ sont notées X_1^M et X_2^M et prennent respectivement des valeurs dans les ensembles de clusters $\{c_{1.}, c_{2.}, ..., c_{k_1.}\}$ et $\{c_{.1}, c_{.2}, ..., c_{.k_2}\}$.

Les points sont répartis dans les $k_1 \times k_2$ cellules de la grille de biclustering. On peut alors déduire les effectifs des clusters et donc la loi empirique de chacune des variables et de chacune des variables-partitions.

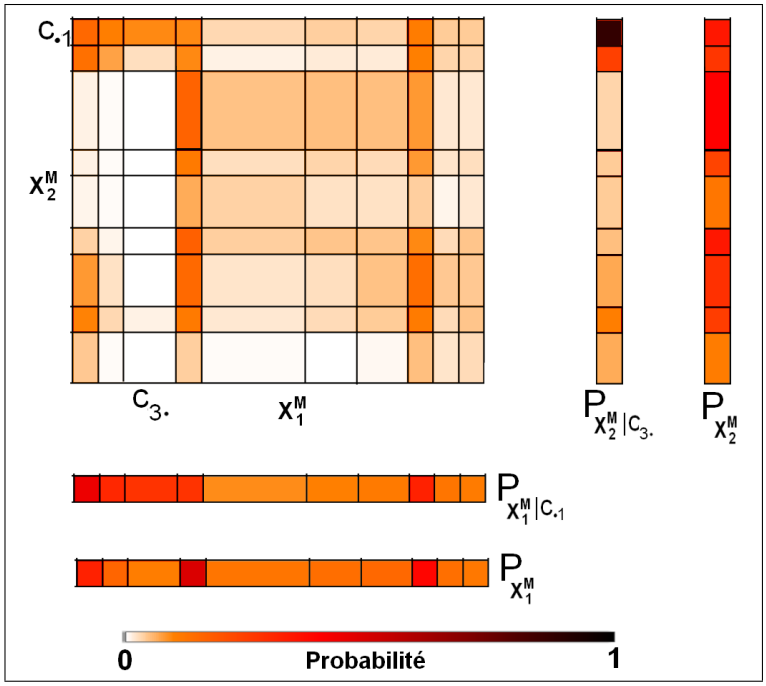


FIGURE 4.1 – Illustration de la loi des biclusters, des variables-partitions et des variables-partitions conditionnellement à un cluster.

On introduit les notations suivantes, illustrées dans la figure 4.1 :

- P_{X_1} (resp. P_{X_2}) est la loi de la variable X_1 (resp. X_2), soit $P_{X_1}(v_{i.}) = P(X_1 = v_{i.})$ (resp. $P_{X_2}(v_{.i}) = P(X_2 = v_{.i})$),
- $P_{X_1^M}$ (resp. $P_{X_2^M}$) est la loi de la variable-partition X_1^M (resp. X_2^M), soit $P_{X_1^M}(c_{i.}) = P(X_1^M = c_{i.})$ (resp. $P_{X_2^M}(c_{.i}) = P(X_2^M = c_{.i})$),
- $P_{X_1^M|v_{.j}}$ (resp. $P_{X_2^M|v_{j.}}$) est la loi de la variable-partition X_1^M (resp. X_2^M) conditionnellement à l'événement $X_2 = v_{.j}$ (resp. $X_1 = v_{j.}$), soit $P_{X_1^M|v_{.j}}(c_{i.}) = P(X_1^M = c_{i.}|X_2 = v_{.j})$ (resp. $P_{X_2^M|v_{i.}}(c_{.i}) = P(X_2^M = c_{.i}|X_1 = v_{i.})$),

– $P_{X_1^M|c_{.j}}$ (resp. $P_{X_2^M|c_{j.}}$) est la loi de la variable-partition X_1^M (resp. X_2^M) conditionnellement à l'événement $X_2^M = c_{.j}$ (resp. $X_1^M = c_{j.}$), soit $P_{X_1^M|c_{.j}}(c_{i.}) = P(X_1^M = c_{i.}|X_2^M = c_{.j})$ (resp. $P_{X_2^M|c_{j.}}(c_{.i}) = P(X_2^M = c_{.j}|X_1^M = c_{i.})$).

Dans la suite de la thèse, la loi d'une variable-partition conditionnellement à un événement est désigné par la loi d'une variable-partition conditionnellement à un cluster ou une valeur, par abus de langage, pour simplifier la terminologie. Ces lois sont utilisées dans la suite du chapitre pour comparer les clusters et les valeurs, au sens de leur effet sur la loi de l'autre variable ou variable-partition.

Afin d'illustrer les outils d'analyse exploratoire introduits dans ce chapitre, nous utilisons une base de données simple à analyser et interpréter : la base *Adult* (Blake et Merz, 1998) de l'UCI (University of California Irvine). Il s'agit d'une base de 48842 individus décrits par quatorze attributs. Parmi ces attributs, nous nous focalisons sur les variables *occupation* (profession) et *education* (niveau d'enseignement). Plus formellement, les caractéristiques des données $\mathcal{D}$ sont les suivantes :

– X_1 est la variable *occupation* avec $n_1 = 14$ valeurs différentes,
– X_2 est la variable *education* avec $n_2 = 16$ valeurs différentes,
– m =48 842 observations.

La structure de biclustering obtenue en utilisant l'approche MODL sur cette base de données est simple à interpréter et illustre bien les différentes notions qui sont introduites par la suite. L'optimisation du critère permet d'obtenir neuf clusters d'*occupation* et dix clusters d'*education*. La grille de biclustering est présentée en figure 4.2 et la composition des clusters dans les tables 4.1 et 4.2.

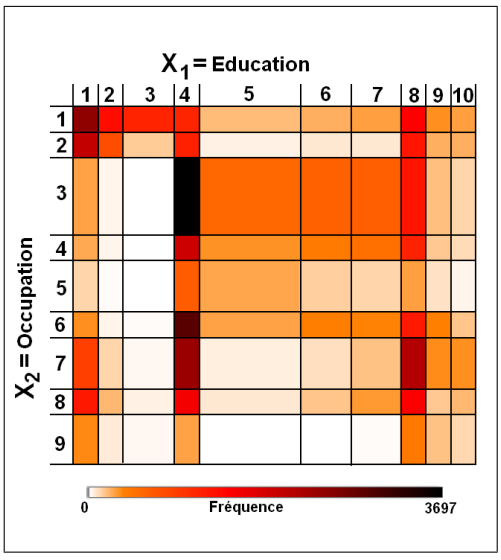


FIGURE 4.2 – Grille de biclustering

Occupation	
Id	Composition
1	Prof-specialty
2	Exec-Managerial
3	Machine-op-inspct
	Transport-moving
	Handlers-cleaners
4	Other-service
5	Farming-fishing
	Priv-house-serv
6	Craft-repair
7	Adm-clerical
	Protective-serv
8	Sales
9	Tech-support
	Armed-Forces

TABLE 4.1 – Partition de la variable *Occupation*

Education	
Id	Composition
1	Bachelors
2	Masters
3	Prof-school
	Doctorate
4	HS-grade
5	Preschool
	1st-4th
	5th-6th
	7th-8th
6	9th / 10th
7	11th / 12th
8	Some-college
9	Assoc-voc
10	Assoc-acdm

TABLE 4.2 – Partition de la variable *Education*

La partition conjointe des variables *education* et *occupation* permet une interprétation duale des résultats : les niveaux d'éducation allant de la maternelle

à la 8*e* classe sont regroupés dans le cluster 5 (voir table 4.2) car les individus ayant arrêté leurs études dans ces classes exercent des professions similaires d'après les données. Les employés administratifs et les forces de l'ordres sont regroupés dans le cluster 7 (voir table 4.1) car les personnes à ces postes ont des niveaux d'instructions similaires. La figure 4.2 permet d'évaluer les relations entre clusters d'*education* et d'*occupation*. Plus la cellule est sombre, plus il y a d'individus dans la base occupant un poste du cluster d'*occupation* avec un niveau d'instruction du cluster d'*education*. Par exemple, peu d'employés administratifs et de forces de l'ordre ont un niveau d'instruction inférieur à la douzième classe (équivalent américain de la terminale), et peu de manutentionnaires, de chauffeurs routiers et d'agents d'entretien (cluster 3) ont poursuivi leurs études au-delà du lycée.

4.2 Rappels de théorie de l'information

Cette partie a pour but de rappeler quelques notions de théorie de l'information et de présenter leurs applications sur les grilles.

4.2.1 L'entropie de Shannon

Shannon (1948) introduit la notion de coût de transmission de l'information. Il s'agit de définir le nombre de bits minimal nécessaire pour transmettre une information : plus il y a d'informations différentes à transmettre, plus le nombre de bits nécessaire pour les transmettre est élevé.

Définition 20. *Soit une variable discrète X prenant n valeurs différentes $\{v_1, v_2, ..., v_n\}$. P est la loi de la variable X, on note $H(X)$ l'entropie de Shannon de la variable X.*

$$H(X) = -\sum_{i=1}^{n} P(v_i) \log P(v_i) \tag{4.1}$$

(a) $P_{X_2^M|c_{.1}}$ (b) $P_{X_1^M|c_{.3}}$

FIGURE 4.3 – $H(X_1^M|c_{.1}) = 1{,}99$ et $H(X_1^M|c_{.3}) = 1{,}58$. Le cluster $c_{.1}$ groupe des professions dans lesquelles les individus ont des niveaux d'instruction plus mélangés que pour le cluster $c_{.3}$.

Dans l'exemple de la figure 4.3, l'entropie de la variable-partition X_1^M conditionnellement au cluster $c_{.1}$ est plus forte que l'entropie de la variable-partition X_1^M conditionnellement au cluster $c_{.3}$. Cela signifie que les individus occupant des professions spécialisées ont des niveaux d'études plus variés que

les personnes occupant des postes de manutentionnaires, de chauffeurs routiers et d'agent d'entretien.

4.2.2 La divergence de Kullback-Leibler

La divergence de Kullback-Liebler (Kullback et Leibler, 1951) est une mesure non-symétrique. On note $KL(P_1||P_2)$ la divergence de Kullback-Leibler de P_1 par rapport à P_2. Cette mesure s'interprète comme le nombre moyen de bits nécessaire pour coder des données tirés suivant la loi P_1, en utilisant un code optimal pour la loi P_2.

Définition 21. *Soient deux lois de probabilités discrètes P_1 et P_2 à valeurs sur $X = \{v_1, v_2, ..., v_n\}$, on note $KL(P_1||P_2)$ la divergence de Kullback-Leibler de P_2 par rapport à P_1.*

$$KL(P_1||P_2) = \sum_{i=1}^{n} P_1(v_i) \log \frac{P_1(v_i)}{P_2(v_i)} \tag{4.2}$$

Bien que cette divergence soit positive ou nulle, elle ne possède pas de borne supérieure et peut donc potentiellement prendre des valeurs infinies. Il suffit que la loi de base prenne une valeur nulle sur un événement du support et que la loi estimée prenne une valeur non-nulle sur le même événement, pour le ratio de P_1 sur P_2 soit infini. Dans ce cas, la divergence de Kullback-Leibler a une valeur infinie.

Reprenons les exemples de lois de la variable-partition X_1^M conditionnellement aux clusters $c_{.1}$ et $c_{.3}$, illustrés dans la figure 4.3 : $KL(P_{X_1^M|c_{.1}}||P_{X_1^M|c_{.3}}) =$ 0,66 et $KL(P_{X_1^M|c_{.3}}||P_{X_1^M|c_{.1}}) =$ 0,37. Il est donc plus coûteux de coder des données tirées suivant la loi $P_{X_1^M|c_{.1}}$ en utilisant un code basé sur la loi $P_{X_1^M|c_{.3}}$ que l'inverse.

4.2.3 La divergence de Jensen-Shannon

La divergence de Jensen-Shannon (Lin, 1991) est une mesure issue d'une somme de divergences de Kullback-Leibler. À l'image de la divergence de Kullback-Leibler, elle n'est pas une distance. Elle présente malgré tout des propriétés intéressantes qui la rendent plus interprétable et plus facile à utiliser que la divergence de Kullback-Leibler.

Définition 22. *Soient deux lois de probabilités discrètes P_1 et P_2, on note $JS^{\alpha_1,\alpha_2}(P_1, P_2)$ la divergence de Jensen-Shannon entre P_1 et P_2 :*

$$JS^{\alpha_1,\alpha_2}(P_1, P_2) = \alpha_1 KL(P_1||\alpha_1 P_1 + \alpha_2 P_2) + \alpha_2 KL(P_2||\alpha_1 P_1 + \alpha_2 P_2) \tag{4.3}$$

avec α_1 et α_2 les coefficients de mélange des lois P_1 et P_2 tels que $\alpha_1, \alpha_2 \in [0, 1]$ et $\alpha_1 + \alpha_2 = 1$. Prendre un coefficient de mélange nul revient à ignorer la loi qui lui est associée.

La divergence de Jensen-Shannon équivaut à moyenner les divergences de Kullback-Liebler du mélange des deux lois étudiées par rapport à chacune d'entre elles. En mélangeant deux lois, si on obtient une nouvelle loi qui estime bien les deux lois desquelles elle est issue, alors la divergence de Jensen-Shannon est faible. Cela signifie qu'en les mélangeant, le nombre moyen de bits supplémentaires nécessaires pour coder chacune d'entre elles à partir de la loi créée est peu élevé.

Cette mesure est symétrique et donc bien plus simple à utiliser pour comparer deux lois que la divergence de Kullback-Leibler. Elle est bornée inférieurement en zéro, comme la divergence de Kullback-Leibler. Pour ce qui est de la borne supérieure, elle est atteinte lorsque pour chaque événement du support, l'une des loi prend une valeur nulle et l'autre non. Dans ce cas, la divergence de Jensen-Shannon vaut $\alpha_1 \log \alpha_1 + \alpha_2 \log \alpha_2$.

Pour résumer, la divergence de Jensen-Shannon est une mesure symétrique, positive et bornée. La divergence est nulle si et seulement si deux lois sont identiques. Toutes ces propriétés en font une semi-métrique.

Théorème 1. *Soient deux lois de probabilités discrètes P_1 et P_2, associées aux coefficients de mélange α_1 et α_2, la divergence de Jensen-Shannon entre deux lois vérifie certaines propriétés des métriques mais n'en est pas une.*

- $JS^{\alpha_1,\alpha_2}(P_1, P_2) \geq 0$,
- $P_1 = P_2 \Rightarrow JS^{\alpha_1,\alpha_2}(P_1, P_2) = 0$,
- $JS^{\alpha_1,\alpha_2}(P_1, P_2) = JS^{\alpha_2,\alpha_1}(P_2, P_1)$.

Démonstration. Les propriétés de métrique de la divergence de Jensen-Shannon ont été étudiées par Lin (1991). □

Jusqu'à présent, on ne s'est intéressé qu'au cas particulier de deux lois associées à des coefficients de mélanges différents. Dans le cas général, il est possible de définir une divergence de Jensen-Shannon entre plusieurs lois.

Définition 23. *Soient k lois de probabilités discrètes $P_1, P_2, ..., P_k$, on note $JS^{\alpha_1,\alpha_2,...,\alpha_k}(P_1, P_2, ..., P_k)$ la divergence de Jensen-Shannon entre les k lois :*

$$JS^{\alpha_1,\alpha_2,...,\alpha_k}(P_1, P_2, ..., P_k) = \sum_{i=1}^{k} \alpha_i KL(P_i || \alpha_1 P_1 + \alpha_2 P_2 + ... + \alpha_k P_k) \quad (4.4)$$

avec α_i les coefficients de mélange des lois tels que $\alpha_i \in [0, 1] \forall i = 1..k_2$ et $\sum_{i=1}^{k_2} \alpha_i = 1$.

Il est possible de décomposer la divergence de Jensen-Shannon généralisée entre k lois en une somme de $k - 1$ divergences entre deux lois, à condition qu'aucun coefficient de mélange soit nul.

Théorème 2. *La divergence de Jensen-Shannon généralisée de k lois peut se décomposer en une somme de divergences de Jensen-Shannon généralisées de*

deux lois :

$$JS^{\alpha_1,\alpha_2,...,\alpha_k}(P_1,P_2,...,P_k) = (1-\alpha_k)JS^{\frac{\alpha_1}{1-\alpha_k},\frac{\alpha_2}{1-\alpha_k},...,\frac{\alpha_{k-1}}{1-\alpha_k}}(P_1,P_2,...P_{k-1}) + JS^{1-\alpha_k,\alpha_k}\left(\sum_{j=1}^{k-1}\frac{\alpha_j}{1-\alpha_k}P_j, P_k\right) \quad (4.5)$$

avec α_i *le coefficient de mélange associé à la loi* P_i*, tel que* $\sum_{i=1}^{k}\alpha_i = 1$ *et* $\alpha_k \in]0,1[$.

Démonstration. La démonstration complète est détaillée en annexe 4.9.1. □

Ainsi la divergence de Jensen-Shannon entre k lois est égale à la somme de la divergence entre les $k-1$ premières lois et de la divergence entre la k^e loi et la moyenne des $k-1$ premières lois. On déduit de ce théorème que la racine carrée de la divergence de Jensen-Shannon respecte le théorème de König-Huygens et donc que cette divergence s'interprète comme une variance entre les lois de probabilités associées aux clusters (Baker et Copson, 1950). Dans le cas d'une grille de co-clustering, on peut ainsi considérer la divergence de Jensen-Shannon entre les lois conditionnelles, associées à chacun des clusters sur une dimension, comme la variance de la partition.

4.3 Simplifier une structure de bi-clustering

Bien que l'approche MODL présente un aspect pratique en ne requérant pas de paramètre utilisateur, il se peut que la partition obtenue soit difficile à interpréter dans sa globalité du fait d'un trop grand nombre de clusters. Le nombre de clusters dans le cas d'une variable nominale est compris entre 1 et n_1 (ou n_2), où n_1 (ou n_2) est le nombre de valeurs prises par la variable X_1 (ou X_2). Ainsi, plus le nombre de modalités de la variable est important, plus le nombre potentiel de clusters est élevé. Asymptotiquement, on peut d'ailleurs obtenir autant de clusters que de modalités sans qu'il s'agisse de sur-apprentissage. C'est pour cette raison qu'il peut être intéressant de simplifier la grille de manière à la rendre plus lisible.

4.3.1 Définition d'une mesure de dissimilarité

Afin de procéder à une simplification de la structure de biclustering, une fusion des clusters est envisagée. Pour choisir la fusion qui dégrade le moins la qualité de la grille de biclustering, une mesure de dissimilarité est introduite. Cette dissimilarité est définie comme l'impact de la fusion de deux clusters sur la valeur du critère optimisé dans l'approche MODL.

Définition 24. *Soient $c_{.1}$ et $c_{.2}$ deux clusters issus de la variable-partition X_2^M selon le modèle $\mathcal{M}$. On note $\mathcal{M}_{(c_{.1}\cup c_{.2})}$ le modèle de biclustering avec les clusters $c_{.1}$ et $c_{.2}$ fusionnés. La dissimilarité $\Delta(c_{.1}, c_{.2})$ entre deux clusters est définie comme la différence du critère MODL après et avant fusion des clusters :*

$$\Delta(c_{.1}, c_{.2}) = \xi(\mathcal{M}_{(c_{.1}\cup c_{.2})}) - \xi(\mathcal{M}) \tag{4.6}$$

Cette mesure de dissimilarité peut s'exprimer comme le logarithme négatif d'un facteur de diminution de la probabilité a posteriori du modèle :

$$P(\mathcal{M}_{(c_{.1}\cup c_{.2})}|\mathcal{D}) = e^{-\Delta(c_{.1},c_{.2})} P(\mathcal{M}|\mathcal{D}) \tag{4.7}$$

Lorsque le nombre d'observations tend vers l'infini, les effets de l'a priori deviennent négligeables devant les termes de vraisemblance du modèle et la dissimilarité entre deux clusters se ramène à une divergence de Jensen-Shannon entre les lois de la variable-partition X_1^M conditionnellement aus clusters $c_{.1}$ et $c_{.2}$.

Théorème 3. *Soient $c_{.1}$ et $c_{.2}$ deux clusters appartenant à la partition X_2^M de la variable X_2. Dans le régime asymptotique, la dissimilarité $\Delta(c_{.1}, c_{.2})$ entre ces deux clusters s'interprète comme la divergence de Jensen-Shannon entre les lois $P_{X_1^M|c_{.1}}$ et $P_{X_1^M|c_{.2}}$ de la variable-partition X_1^M, conditionnellement aux clusters $c_{.1}$ et $c_{.2}$:*

$$\lim_{m\to+\infty} \frac{\Delta(c_{.1}, c_{.2})}{m} = (P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2}))JS^{\alpha_1,\alpha_2}(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}}) \tag{4.8}$$

$$\text{avec } \alpha_1 = \frac{P_{X_2^M}(c_{.1})}{P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2})} \text{ et } \alpha_2 = \frac{P_{X_2^M}(c_{.2})}{P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2})}$$

Démonstration. En calculant la variation de la vraisemblance liée à la fusion de deux clusters et en appliquant l'approximation de Stirling : $\log(m!) = m\log(m)$ lorsque $m \to +\infty$ (Abramowitz et Stegun, 1965), on arrive à ce résultat. La démonstration est détaillée en annexe 4.9.2. □

La définition du coût de fusion a été introduit sur la variable X_2 mais la même démarche peut être menée sur la variable X_1.

On a vu précédemment que la divergence de Jensen-Shannon possède certaines propriétés des métriques. Cette divergence compare deux lois alors que la mesure de dissimilarité définie ici compare deux clusters auxquels sont associés une loi et un coefficient de mélange. Ainsi, si deux clusters sont identiques, cela implique que leur dissimilarité est nulle mais la réciproque n'est pas vérifiée : deux clusters différents peuvent avoir une dissimilarité nulle si les lois de probabilité de X_1^M conditionnellement à ces clusters sont identiques. La mesure de dissimilarité entre deux clusters possède malgré tout les propriétés nécessaires des mesures de dissimilarité.

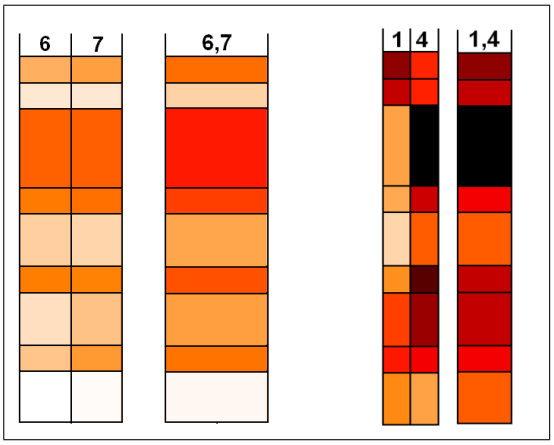


FIGURE 4.4 – Fusion des clusters 6 and 7 (la moins coûteuse) et des clusters 1 et 4 (la plus coûteuse) pour la variable *education*

Afin d'illustrer ces propriétés, les résultats du biclustering sur la base *Adult* sont analysés. En calculant la dissimilarité entre chaque paire de clusters de la partition de la variable *education*, il apparaît que les clusters les plus similaires sont les clusters 6 (9^e et 10^e classes, équivalents américains de la 3^e et de la 2^{nde}) et 7 (11^e et 12^e classes, équivalents de la 1^{re} et de la terminale) alors que les clusters 1 (Bachelor, équivalent de la Licence) et 4 (High School grade, équivalent du lycée) sont les plus différents. En observant la figure 4.4, on remarque que la loi de la variable-partition X_2^M (*Occupation*) conditionnellement au nouveau cluster – obtenu par fusion des clusters 6 et 7 – est très similaire à la loi de X_2^M conditionnellement aux clusters dont il est issu. En d'autres termes, les personnes ayant des niveaux d'études compris entre la troisième et la terminale occupent des postes similaires. La valeur exacte de cette dissimilarité est de 1,71, ce qui est très faible ($8{,}4 \times 10^{-4}\%$ de la valeur totale du critère). L'estimation de la dissimilarité (c'est-à-dire la valeur sous l'hypothèse asymptotique) est de 27,32. Quant à la paire de clusters les plus différents, on observe sur la figure 4.4 que le cluster issu de la fusion des deux clusters est très différent des clusters desquels il provient. La valeur exacte de la dissimilarité est de 3275,40, et la valeur estimée, 3293,11, ce qui correspond environ à 1,6% de la valeur du critère. Dans le cas de deux clusters très similaires, l'erreur entre l'estimation et la vraie valeur s'explique par les termes d'a priori dont la variation est conséquente par rapport à celle de la vraisemblance. Dans le cas inverse, l'erreur est moins importante et l'estimation correcte. Notons que la différence entre valeurs exacte et estimée est quasiment constante. En effet, la plupart des termes de l'a priori sont indépendants du choix des clusters qu'on fusionne. Dans un soucis de précision et de cohérence avec l'approche, on préfère utiliser la variation exacte du critère plutôt que son approximation asymptotique.

4.3.2 Classification hiérarchique ascendante

Maintenant qu'une mesure de dissimilarité a été définie, un post-traitement consistant à simplifier le modèle peut être introduit. Pour ce faire, une classification hiérarchique ascendante est proposée. Il s'agit de fusionner successivement les clusters depuis le niveau le plus fin, jusqu'à obtenir un seul cluster. Vu que la dissimilarité est directement déduite du critère optimisé pour construire la grille de biclustering, ce post-traitement doit être considéré comme un outil d'analyse exploratoire de la structure. Afin d'avoir une visualisation appropriée pour une analyse exploratoire, nous proposons de construire un dendrogramme et une courbe de Pareto cumulative. Le dendrogramme est un arbre présentant la hiérarchie des clusters obtenus par fusions successives. Quant à la courbe de Pareto, on trace la valeur du critère de l'approche MODL, en fonction du nombre de clusters conservés dans le modèle.

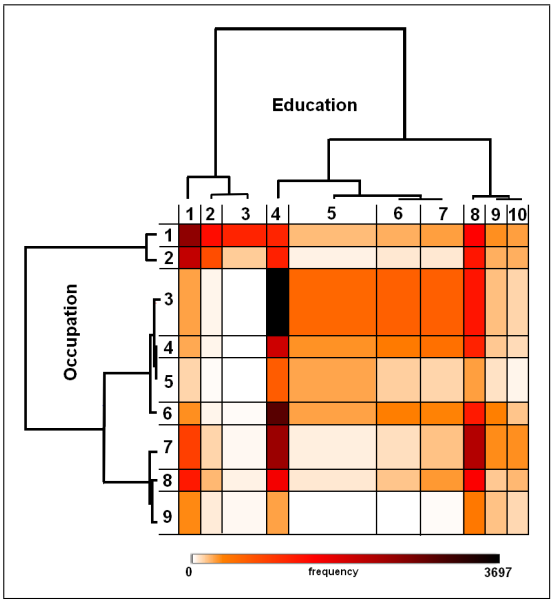
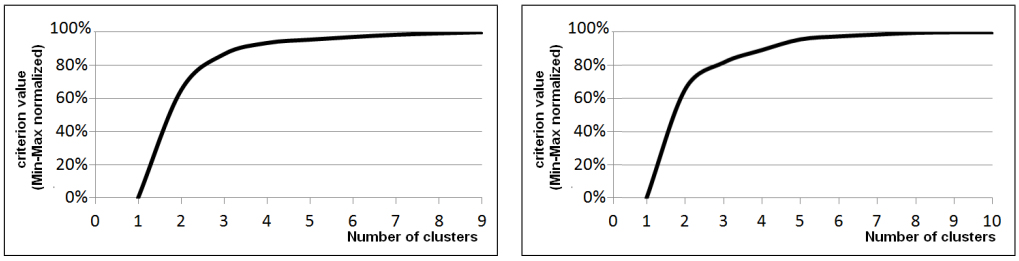


FIGURE 4.5 – Dendrogrammes pour les partitions des variables *education* et *occupation*

À chaque niveau de la hiérarchie on définit le taux d'information du modèle comme le pourcentage de la valeur optimale du critère MODL conservée après plusieurs fusions, par rapport au modèle ne contenant qu'un seul cluster.

Définition 25. *Soient un modèle $\mathcal{M}$, le modèle optimal selon MODL $\mathcal{M}^*$ et le modèle nul (avec un seul co-cluster) $\mathcal{M}_\emptyset$. Le taux d'information du modèle $\mathcal{M}$ est défini comme :*

$$\tau(\mathcal{M}) = \frac{\xi(\mathcal{M}) - \xi(\mathcal{M}_\emptyset)}{\xi(\mathcal{M}^*) - \xi(\mathcal{M}_\emptyset)} \tag{4.9}$$



(a) Variable *education* (b) Variable *occupation*

FIGURE 4.6 – Courbes de Pareto cumulatives pour les partitions des variables *education* et *occupation*

La figure 4.5 présente les dendrogrammes pour les partitions des variables *education* et *occupation*. Nous remarquons, en particulier pour la variable *occupation*, que les premières fusions sont beaucoup moins coûteuses que les dernières. En effet, en analysant la courbe de Pareto cumulative pour cette partition (voir la figure 4.6b), nous observons qu'avec seulement trois clusters, la taux d'information du modèle est de 82%. Après analyse des dendrogrammes et des courbes de Pareto, nous choisissons de conserver trois clusters pour analyser la partition de la variable *occupation*. Cette partition simplifiée est alors composée des trois groupes suivant : {*prof-speciality, exec-managerial*}, {*machine-op-inspct, transport-moving, handlers-cleaners, other service, farming-fishing, priv-house-serv, craft-repair*} et {*adm-clerical, protective-serv, sales, tech-support, armed-forces*}. Une analyse similaire peut être faite sur la partition de la variable *education*.

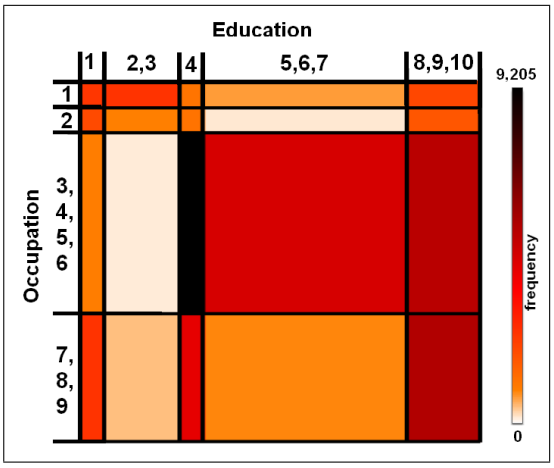


FIGURE 4.7 – Grille de biclustering simplifiée

Dans certains cas, on peut vouloir étudier les deux partitions simplifiées conjointement. Les fusions sont faites séquentiellement plutôt qu'indépendamment sur chaque partition. À chaque étape, la meilleure fusion parmi toutes les

fusions possibles sur les deux partitions est effectuée. L'ordre des fusions des clusters peut différer de celui qu'il aurait été si seulement une partition avait été simplifiée.

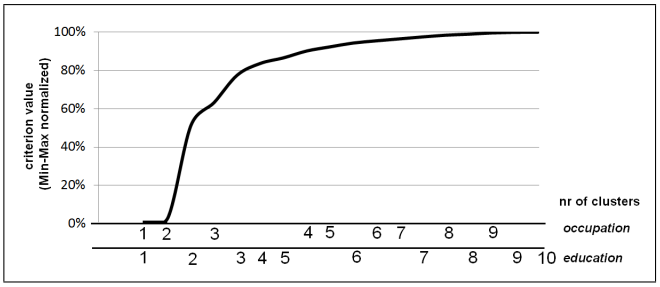


FIGURE 4.8 – Courbe cumulative de Pareto pour les deux partitions

Les figures 4.7 et 4.8 montrent qu'en réduisant le nombre de clusters de la partition de la variable *occupation* de neuf à quatre et le nombre de clusters de la partition de la variable *education* de dix à cinq, le taux d'information du modèle est de 90%.

Remarquons que ce post-traitement est très proche des travaux de Slonim et Tishby (2000) où la divergence de Jensen-Shannon est directement utilisée pour construire la hiérarchie des clusters sur chacune des variables. L'approche MODL, à la différence de l'approche de Slonim et Tishby (2000), utilise un critère régularisé. Ainsi, lorsque les données sont en nombre suffisant pour que les hypothèses asymptotiques soient valides, l'approche MODL ne présente pas d'intérêt par rapport à l'approche de Slonim et Tishby (2000). En effet, dans un tel cas, on obtient avec MODL autant de clusters que de valeurs au niveau le plus fin pour chaque variable, et la mesure de dissimilarité utilisée pour le post-traitement est proportionnelle à la mesure de Slonim et Tishby (2000). Ainsi, les deux approches produisent les mêmes dendrogrammes. Cependant, lorsque les données ne sont pas assez nombreuses, l'approche MODL produit un niveau le plus fin de clustering, ce qui évite à l'utilisateur de choisir un niveau de grain plus précis mais avec des clusters non significatifs. Enfin, lorsque les données sont aléatoires, l'approche MODL produit un seul cluster sur chaque variable, ce qui évite à l'utilisateur de trouver une structure qui n'existe pas.

Le régime asymptotique est atteint plus ou moins rapidement en fonction de la complexité des structures à détecter. On voit, par exemple, que dans les graphes avec des schémas en quasi-cliques (voir section 3.2.4), la détection de la structure peut demander un nombre d'observations (arêtes) de l'ordre du nombre de nœuds au carré. Il en faut beaucoup moins pour les schémas purs ($p = 1$) et beaucoup plus pour les schémas bruités ($p = 0{,}6$). Si on transpose cette observation au cas général des variables nominales, on s'attend à ce qu'une structure relativement bruitée soit détectée avec un nombre d'observations de l'ordre du produit des nombres de modalités. Si la structure sous-jacente est très proche du cas d'indépendance entre les variables, il faut beaucoup plus

d'observations, alors que dans le cas d'une dépendance très forte des variables, on a besoin de peu d'observations pour faire émerger les structures.

4.4 Notions d'inertie dans le biclustering

Dans les problèmes de clustering, calculer l'inertie est une méthode permettant d'évaluer la qualité de la partition (voir le chapitre 2). Dans de nombreuses approches, l'inertie est directement optimisée de manière à obtenir une bonne segmentation des données, comme par exemple dans les k-means ou les classifications hiérarchiques ascendantes avec un lien de Ward (Duda *et al.*, 2001). Il existe trois types d'inerties : l'inertie inter-clusters, intra-cluster et l'inertie totale. Nous proposons dans cette section de définir ces notions pour les grilles de biclustering.

4.4.1 Inertie inter-clusters

L'inertie inter-clusters est une mesure qui permet de quantifier la dispersion des clusters. Elle peut s'interpréter comme une variance des groupes. Dans tous les problèmes de clustering simple, cette mesure est maximisée. En effet le but d'un clustering est d'obtenir une représentation simplifiée des données telle que tous les clusters soient les moins similaires les uns par rapport aux autres. Dans le cas de l'approche MODL, l'optimisation du critère peut s'interpréter comme une maximisation de la dissimilarité moyenne entre les clusters. La dissimilarité moyenne est définie comme la variation du critère due à la fusion de l'ensemble des clusters en un unique cluster, sur une seule des partitions.

Définition 26. *Soit X_2^M, variable-partition suivant les paramètres d'un modèle $\mathcal{M}$. On note $\mathcal{M}_{X_1^M \emptyset}$ le modèle tel que les k_2 clusters issus de la variable-partition X_2^M ont été fusionnés en un seul cluster. L'inertie inter-clusters J_{inter} de X_2^M est alors définie de la manière suivante :*

$$J_{inter}(X_2^M) = \xi(\mathcal{M}_{X_1^M \emptyset}) - \xi(\mathcal{M}) \tag{4.10}$$

L'inertie inter-clusters peut se décomposer en un terme constant et une somme sur les clusters de la partition étudiée.

$$\begin{aligned} J_{inter}(X_2^M) =& f(m, n_1, n_2, k_1, k_2) - \sum_{j=1}^{k_2} \log \binom{m_{.j}^C + n_{.j}^C - 1}{n_{.j}^C - 1} \\ &+ \log \frac{m!}{\prod_{i=1}^{k_1} m_{i.}^C!} - \sum_{j=1}^{k_2} \log \frac{m_{.j}^C!}{\prod_{i=1}^{k_1} m_{ij}^C!} \end{aligned} \tag{4.11}$$

où f est une fonction indépendante du contenu des clusters de la partition X_2^M.

De manière similaire à la définition 24, l'inertie inter-clusters peut s'interpréter comme un facteur de perte de probabilité du modèle de biclustering lié à la fusion des clusters d'une partition en un seul cluster. Asymptotiquement, l'inertie inter-clusters est proportionnelle à une divergence de Jensen-Shannon entre les lois de X_1^M conditionnellement à tous les clusters de la variable-partition X_2^M.

Théorème 4. *L'inertie inter-clusters de la variable-partition X_2^M converge asymptotiquement vers la divergence de Jensen-Shannon entre les lois $P_{X_1^M|c_{.1}}$, $P_{X_1^M|c_{.2}}$, ... et $P_{X_1^M|c_{.k_2}}$.*

$$\begin{aligned}\lim_{m\to+\infty} \frac{J_{inter}(X_2^M)}{m} &= JS^{\alpha_1,\alpha_2,...,\alpha_{k_2}}(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}}, ..., P_{X_1^M|c_{.k_2}}) \\ &= \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) KL(P_{X_1^M|c_{.j}} || P_{X_1^M})\end{aligned} \tag{4.12}$$

avec $\alpha_i = P_{X_2^M}(c_{.i})$.

Démonstration. La démonstration est la même que la démonstration de la limite de la dissimilarité (voir théorème 3). □

La divergence de Jensen-Shannon est bornée, donc l'inertie inter-clusters l'est également. Afin d'en avoir une meilleure interprétation, on étudie les bornes de cette mesure. La valeur minimale de l'inertie est zéro : elle caractérise une partition dont tous les clusters suivent la même loi de probabilité. Ce cas, en pratique, ne peut être pas observé car l'approche MODL est régularisée et ne fait pas de sur-partitionnement. La borne supérieure de l'inertie inter-clusters correspond au cas où 100% des points sont sur les biclusters diagonaux de la grille de biclustering.

Théorème 5. *Asymptotiquement, l'inertie inter-clusters J_{inter} est une mesure bornée :*

$$\lim_{m\to+\infty} \frac{J_{inter}(X_2^M)}{m} \in [0, H(X_2^M)] \tag{4.13}$$

Démonstration. De la même manière que la divergence de Jensen-Shannon admet une borne inférieure en 0, la valeur asymptotique de l'inertie inter-clusters est positive ou nulle. Pour la borne supérieure, on a vu dans la section 4.2 que la divergence de Jensen-Shannon admet une borne supérieure $\alpha_{.1} \log \alpha_{.1} + \alpha_{.2} \log \alpha_{.2} + ... + \alpha_{.k_2} \log \alpha_{.k_2}$ avec $\alpha_{.i}$ le coefficient de mélange associé à la $i^{ème}$ loi conditionnelle. Ici tous les clusters sont mélangés et le coefficient de mélange correspond à la probabilité pour une observation d'appartenir à un cluster : $\alpha_{.i} = P_{X_2^M}(c_{.i})$. La borne supérieure est donc égale à $P_{X_2^M}(c_{.1}) \log P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2}) \log P_{X_2^M}(c_{.2}) + ... + P_{X_2^M}(c_{.k_2}) \log P_{X_2^M}(c_{.k_2}) = H(X_2^M)$.

□

Finalement, l'inertie inter-clusters est une quantification de la corrélation entre les variables-partitions : plus une variable-partition peut expliquer l'autre variable-partition, plus la valeur de l'inertie est forte, le cas maximal étant le cas d'une grille de biclustering diagonal où chaque cluster de X_1^M interagit avec un unique cluster de X_2^M. Ainsi, l'inertie inter-clusters peut s'interpréter comme l'information mutuelle entre les deux variables-partitions.

Théorème 6. *L'inertie inter-cluster J_{inter} est asymptotiquement proportionnelle à l'information mutuelle entre les deux variables-partitions X_1^M et X_2^M :*

$$\lim_{m\to+\infty} \frac{J_{inter}(X_2^M)}{m} = MI(X_1^M; X_2^M) \tag{4.14}$$

Démonstration. La démonstration est détaillée en annexe 4.9.3. □

Pour un co-clustering de deux variables, l'inertie inter-clusters de la partition de la première variable est la même que l'inertie inter-clusters de la partition de la seconde variable. L'inertie inter-clusters de la partition X_2^M se calcule pour une partition fixée X_1^M de la variable X_1. L'inertie inter-clusters mesure la dispersion des clusters de la partition X_2^M, ce qui peut s'interpréter comme l'information apportée par la partition X_2^M conditionnellement à X_1^M. Pour une meilleure interprétation des résultats, l'inertie inter-clusters peut être normalisée par son maximum.

Définition 27. *Pour une variable nominale X_2, le taux de dispersion des clusters de sa partition X_2^M est défini comme l'inertie inter-clusters normalisée :*

$$\tau_{inter}(X_2^M) = \frac{J_{inter}(X_2^M)}{mH(X_2^M)} \tag{4.15}$$

Asymptotiquement,

$$\begin{aligned}\lim_{m\to+\infty} \tau_{inter}(X_2^M) &= \frac{JS^{\alpha_1,\alpha_2,\ldots,\alpha_{k_2}}(P_{X_1^M|c.1}, P_{X_1^M|c.2}, \ldots, P_{X_1^M|c.k_2})}{H(X_2^M)} \\ &= \frac{MI(X_1^M; X_2^M)}{H(X_2^M)}\end{aligned} \tag{4.16}$$

Dans le cas de deux variables nominales, cette définition du taux de dispersion est très proche de la définition de l'information mutuelle normalisée (Strehl et Ghosh, 2003) et plus précisément de la variante connue sous le nom de coefficient de contrainte (Coombs *et al.*, 1970), information mutuelle normalisée par l'entropie d'une des deux variable étudiées, qui est utilisée pour estimer la qualité d'une partition et d'un co-clustering de manière générale.

Dans le cas de la base *adult*, le taux de dispersion de la partition de la variable *education* vaut 9,09% (9,35% avec l'approximation asymptotique) alors que le taux de dispersion de la partition de la variable *occupation* vaut 8,32% (8,57% avec l'approximation asymptotique). Cela signifie que les clusters de la partition de la variable *education* sont moins similaires les uns par rapport aux autres que les clusters de la partition de la variable *occupation*. On note que les valeurs des taux de dispersion sont relativement faibles, caractéristique d'une structure de biclustering non-diagonale.

4.4.2 Inertie intra-cluster

L'inertie intra-cluster est définie de manière analogue à l'inertie inter-clusters. Au lieu d'étudier la dispersion des clusters, on se focalise sur la dispersion des valeurs au sein de chaque cluster. Pour ce faire, on calcule le coût de division d'un cluster en autant de clusters qu'il regroupe de valeurs. Plus les éléments sont dispersés, plus l'inertie intra-cluster est grande.

Définition 28. *Soit X_2^M, la variable-partition suivant les paramètres d'un modèle $\mathcal{M}$. On note $\mathcal{M}_{X_1^M X_2}$ le modèle tel que tous les k_2 clusters issus de la variable-partition X_2^M ont été divisés en n_2 nouveaux clusters, ce qui équivaut à considérer la variable-partition X_2^M comme la variable X_2 elle-même. On a donc autant de clusters dans X_2^M que de valeurs dans X_2. L'inertie intra-cluster J_{intra} de X_2^M est alors définie de la manière suivante :*

$$J_{intra}(X_2^M) = \xi(\mathcal{M}) - \xi(\mathcal{M}_{X_1^M X_2}) \tag{4.17}$$

L'inertie intra-cluster peut se décomposer en un terme constant et une somme sur les clusters de la partition étudiée.

$$\begin{aligned} J_{intra}(X_2^M) =& f(m, n_1, n_2, k_1, k_2) + \sum_{j=1}^{k_2} \log \binom{m_{.j}^C + n_{.j}^C - 1}{n_{.j}^C - 1} \\ &+ \sum_{j=1}^{k_2} \log \frac{m_{.j}^C!}{\prod_{i=1}^{k_1} m_{ij}^C!} - \sum_{j=1}^{n_2} \log \frac{m_{.j}!}{\prod_{i=1}^{k_1} m_{ij}!} \end{aligned} \tag{4.18}$$

De manière similaire à l'inertie inter-clusters, l'inertie intra-cluster s'interprète asymptotiquement comme une somme pondérée de divergences de Jensen-Shannon entre les lois de la variable-partition X_1^M conditionnellement à toutes les valeurs de la variable X_2.

Théorème 7. *L'inertie intra-cluster de la variable-partition X_2^M converge asymptotiquement vers une somme pondérée de divergences de Jensen-Shannon*

entre les lois $P_{X_1^M|v_{.1}}$, $P_{X_1^M|v_{.2}}$, ... *et* $P_{X_1^M|v_{.n_2}}$.

$$\begin{aligned}\lim_{m\to+\infty}\frac{J_{intra}(X_2^M)}{m} &= \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j})JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i}\in c_{.j}\}) \\ &= \sum_{i=1}^{n_2} P_{X_2}(v_{.i}) \underset{v_{.i}\in c_{.j}}{KL}(P_{X_1^M|v_{.i}}||P_{X_1^M|c_{.j}})\end{aligned} \tag{4.19}$$

Notons que non-asymptotiquement, l'inertie intra-cluster est négative et positive asymptotiquement. Dans le premier cas, le modèle optimal n'est pas le modèle le plus fin, ce qui explique que l'inertie intra-cluster est négative : les données ne sont pas en nombre suffisant pour que la vraisemblance domine le coût de codage du modèle le plus fin. Asymptotiquement, les données sont en quantité suffisante pour que chaque valeur soit distinguée des autres et dans ce cas le modèle optimal est le modèle le plus fin. En se plaçant dans le cadre asymptotique, un modèle avec des partitions ayant un nombre réduit de clusters n'est plus optimal contrairement au modèle le plus fin. Ainsi, le modèle le plus fin est moins coûteux que le modèle étudié et l'inertie intra-cluster devient positive.

Les même propriétés asymptotiques sont observées pour l'inertie intra- et inter- clusters.

Théorème 8. *L'inertie intra-cluster* J_{intra} *est une mesure bornée :*

$$\lim_{m\to+\infty}\frac{J_{intra}(X_2^M)}{m} \in \left[0, \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j})H(X_2|c_{.j})\right] \tag{4.20}$$

Démonstration. La démonstration est similaire à celle du théorème 5 □

Théorème 9. *L'inertie intra-cluster* J_{intra} *est asymptotiquement proportionnelle à la somme des informations mutuelles entre la variable-partition* X_1^M *et la variable* X_2 *conditionnellement à chaque cluster :*

$$\lim_{m\to+\infty}\frac{J_{intra}(X_2^M)}{m} = \sum_{j=1}^{k_2} MI(X_1^M; X_2|c_{.j}) \tag{4.21}$$

Démonstration. La démonstration est similaire à celle du théorème 6 □

De manière similaire à précédemment, le taux de dispersion intra-cluster est défini comme l'inertie intra-cluster normalisée.

Définition 29. *Le taux de dispersion des valeurs au sein de la partition X_2^M est défini comme l'inertie intra-cluster normalisée :*

$$\tau_{inter}(X_2^M) = \frac{J_{intra}(X_2^M)}{\sum_{j=1}^{k_2} m_{.j}^c H(X_2|c_{.j})} \tag{4.22}$$

Asymptotiquement,

$$\begin{aligned}\lim_{m\to+\infty} \tau_{intra}(X_2^M) &= \frac{\sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.j}\})}{\sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) H(X_2|c_{.j})} \\ &= \frac{\sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) MI(X_1^M; X_2|c_{.j})}{\sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) H(X_2|c_{.j})}\end{aligned} \tag{4.23}$$

Le calcul de l'inertie intra-cluster sur les partitions de la base de données *Adult* montre de grandes divergences entre les valeurs asymptotiques et non-asymptotiques : le taux de dispersion des valeurs au sein de la partition de la variable *occupation* vaut $-0{,}90\%$ (23,59 % avec l'approximation asymptotique), celle de la variable *education* vaut $-2{,}54\%$ (5,09 % avec l'approximation asymptotique). Les clusters de la variable *education* sont donc composées de valeurs moins dispersées que pour la variable *occupation*. Afin d'avoir une meilleure interprétation de cette mesure, une utilisation de l'approximation asymptotique est préférable : la variation des termes d'a priori ne sont pas liés à la qualité d'une partition contrairement à la variation des termes de vraisemblance.

4.4.3 Inertie totale

L'inertie totale est un critère global relatif à une partition. Il est composé des inerties inter- et intra- clusters.

Définition 30. *L'inertie totale de la partition X_2^M est définie comme la somme de ses inerties inter- et intra clusters.*

$$\begin{aligned}J(X_2^M) &= J_{inter}(X_2^M) + J_{intra}(X_2^M) \\ &= \xi(\mathcal{M}_{X_1^M\emptyset}) - \xi(\mathcal{M}_{X_1^M X_2})\end{aligned} \tag{4.24}$$

L'inertie totale peut se décomposer en un terme constant et une somme sur les clusters de la partition étudiée.

$$J(X_2^M) = f(m, n_1, n_2, k_1, k_2) + \log \frac{m!}{\prod_{i=1}^{k_1} m_{i.}^C!} - \sum_{j=1}^{n_2} \log \frac{m_{.j}!}{\prod_{i=1}^{k_1} m_{ij}!} \quad (4.25)$$

où f est une fonction indépendante du contenu des clusters de la partition X_2^M.

L'inertie totale correspond à la différence entre le coût du modèle le plus fin et le coût du modèle nul et ne dépend donc que des données et de la partition de l'autre variable. Ainsi en maximisant l'inertie inter-clusters, l'inertie intra-cluster est minimisée.

Théorème 10. *Asymptotiquement, l'inertie totale d'une partition s'interprète comme une divergence de Jensen-Shannon ou comme une information mutuelle.*

$$\begin{aligned} \lim_{m \to +\infty} \frac{J(X_2^M)}{m} &= JS(P_{X_1^M|v.1}, P_{X_1^M|v.2}, ..., P_{X_1^M|v.n_2}) \\ &= \sum_{j=1}^{n_2} P_{X_2}(v_{.j}) KL(P_{X_1^M|v_{.j}} || P_{X_1^M}) \\ &= MI(X_1^M; X_2) \end{aligned} \quad (4.26)$$

Démonstration. voir La démonstration est similaire à celle du théorème 6 □

Pour résumer, l'inertie totale d'une partition a été déduite du critère MODL défini en section 2.2.5 et définie comme la somme des gains obtenu en segmentant une variable en clusters (inertie inter-clusters) et en groupant des valeurs (inertie intra-cluster). Cette mesure s'interprète comme l'information mutuelle entre la variable étudiée et la variable-partition issue de l'autre variable. Cette mesure est donc indépendante de la partition étudiée et ne reflète que sa capacité à être segmentée en clusters connaissant la partition de l'autre variable. Sachant cette propriété, on en déduit que le meilleur modèle est celui qui maximise l'inertie inter-clusters, ce qui est équivalent à minimiser l'inertie intra-cluster. Ce résultat est en accord avec la propriété de Huygens que respecte la mesure de dissimilarité déduite du critère de l'approche MODL et étudié dans la section 4.3.1.

4.5 L'intérêt et la typicité

Dans les problèmes de biclustering et plus particulièrement lorsque le nombre de clusters est important, on peut être confronté à des difficultés d'interprétation des résultats. On a vu précédemment qu'il est possible de procéder à un post-traitement permettant de simplifier la structure étudiée en réduisant le nombre de clusters de façon maîtrisée. Cependant, cette simplification de la partition induit une complexification des clusters : un plus grand nombre de valeurs y sont groupées, il est donc plus difficile de faire une étude locale à chacun des

clusters. Dans l'étude des clusters d'une variable-partition, nous exploitons l'autre variable-partition pour identifier les clusters remarquables et les valeurs représentatives de leur cluster, en analysant leur contribution au critère optimisé pour inférer la structure de biclustering. Ainsi, nous n'avons pas besoin de faire d'hypothèses de dissimilarité dans le cas des variables nominales.

Dans cette partie, on propose de mettre en place des indicateurs permettant, dans un premier temps, de caractériser les clusters les plus intéressants à étudier ; et dans un second temps, de détecter les valeurs les plus représentatives d'un cluster. L'intérêt est donc utile lorsque le nombre de clusters est important et la typicité lorsque le nombre de valeurs par cluster est important.

4.5.1 L'intérêt d'un cluster

L'intérêt des clusters est une mesure qui est particulièrement utile lorsque le nombre de clusters est grand. Dans ce cas, il est parfois intéressant de conserver le niveau de partitionnement le plus fin et d'étudier les clusters qui jouent un rôle significatif dans la structure de biclustering. On fait ce type d'analyse lorsqu'on étudie les phénomènes de niches. On a défini précédemment les notions d'inertie. L'inertie inter-clusters quantifie la dispersion des clusters au sein d'une partition. On définit donc un cluster jouant un rôle significatif comme un cluster ayant un fort impact sur l'inertie inter-clusters.

Définition 31. *Pour un cluster $c_{.\gamma}$ de la variable-partition X_2^M, l'intérêt T_c est défini comme l'impact du cluster sur l'inertie inter-clusters de la partition dont il provient.*

$$J_{inter}(X_2^M) = \sum_{i_2=1}^{k_2} T_c(c_{.i_2}) \tag{4.27}$$

ce qui équivaut à définir l'intérêt comme :

$$\begin{aligned} T_c(c_{.\gamma}) &= J_{inter}(X_2^M) - J_{inter}(X_2^M \setminus c_{.\gamma}) \\ &= (\xi(\mathcal{M}) - \xi(\mathcal{M}_{X_1^M\emptyset})) - (\xi(\mathcal{M}|X_2^M \setminus c_{.\gamma}) - \xi(\mathcal{M}_{X_1^M\emptyset}|X_2^M \setminus c_{.\gamma})) \\ &= P_{X_2^M}(c_{.\gamma}) f(m, n_1, n_2, k_1, k_2) - \log \binom{m_{.\gamma}^C + n_{.\gamma}^C - 1}{n_{.\gamma}^C - 1} \\ &+ P_{X_2^M}(c_{.\gamma}) \log \frac{m!}{\prod_{i=1}^{k_1} m_{i.}^C!} - \log \frac{m_{.\gamma}^C!}{\prod_{i=1}^{k_1} m_{i\gamma}^C!} \end{aligned} \tag{4.28}$$

où $X_2^M \setminus c_{.\gamma}$ est la partition X_2^M de laquelle on a retiré le cluster $c_{.\gamma}$.

Asymptotiquement, l'inertie inter-clusters d'une partition s'interprète comme une divergence de Jensen-Shannon, qui se décompose en une somme de divergences de Kullback-Leibler sur chacun des clusters. On en déduit donc l'estimation asymptotique de la contribution de chacun des clusters à l'inertie inter-clusters de leur partition.

Théorème 11. *Asymptotiquement, l'intérêt d'un cluster $c_{.\gamma}$ de la partition X_2^M s'interprète comme la divergence de Kullback-Leibler de $P_{X_1^M}$ la loi de la variable-partition X_1^M par rapport à $P_{X_1^M|c_{.\gamma}}$ la loi de la variable-partition X_1^M conditionnellement au cluster $c_{.\gamma}$, le tout pondéré par $P_{X_2^M}(c_{.\gamma})$, la probabilité qu'une observation appartienne au cluster $c_{.\gamma}$.*

$$\lim_{m \to +\infty} \frac{T_c(c_{.\gamma})}{m} = P_{X_2^M}(c_{.\gamma}) KL(P_{X_1^M|c_{.\gamma}} || P_{X_1^M}) \tag{4.29}$$

Démonstration. La démonstration est similaire à celle du théorème 6 □

L'intérêt ainsi défini est une mesure permettant de se focaliser sur les clusters intéressants dans le sens où, d'un côté, les lois conditionnelles qui leur sont associées divergent de la loi de la variable-partition (terme de divergence de probabilités), et de l'autre, où un nombre significatif d'observations y prennent valeurs (terme de pondération).

Id	Composition des clusters	intérêt
1	{Prof-specialty}	1
3	{Machine-op-inspct, Transp.-moving, Handl.-cleaners}	0,5095
2	{Exec-managerial}	0,4323
6	{Craft-repair}	0,2514
4	{Other-service}	0,2248
7	{Adm-clerical, Protective-serv}	0,1901
5	{Farming-fishing, Priv-house-serv}	0,1001
9	{Tech-support, Armed-Forces}	0,0946
8	{Sales}	0,0906

TABLE 4.3 – intérêts normalisés des clusters de la partition de la variable *Occupation*

Nous illustrons cette mesure en étudiant la partition de la variable *occupation* de la base de données *Adult*. Le taux de dispersion (inertie normalisée) inter-clusters de cette partition est de 8,32%. L'intérêt des clusters est détaillée dans la table 4.3. La mesure a été normalisée de manière à ce que son maximum soit égal à 1. Ainsi, on peut facilement comparer les contributions de chacun des clusters à l'inertie. Les clusters avec le plus fort intérêt correspondent aux emplois spécialisés ainsi qu'aux emplois non qualifiés avec respectivement des valeurs d'intérêt de 1 et 0,5095. Ces catégories professionnelles se caractérisent par des niveaux de formation spécifiques et représentent une partie importante des observations de la base (resp. 18,39% et 15,25% des observations). Au contraire, les intérêts les plus faibles concernent les professions du commerce ainsi que des supports techniques/forces armés avec des intérêts respectifs de 0,0906 et 0,0946. Les professions du cluster supports techniques/forces armés

représentent moins de 3% des individus de la base de données. Ce cluster a donc un faible intérêt car il groupe des professions dont les niveaux d'études ne sont pas suffisamment spécifiques pour contrer leur sous-représentation. Dans le cas des professions liées au commerce, elles sont occupées par 11,27% des individus de la base de données, ce qui en fait la cinquième catégorie professionnelle la plus représentée sur les quatorze de la base. La faible valeur d'intérêt de cette catégorie est liée au fait que les personnes occupant des postes dans le commerce ont des niveaux de formation similaires à l'ensemble des individus de la base.

4.5.2 Typicité d'une valeur

On s'intéresse ici à la détection des valeurs les plus représentatives de chaque cluster. Pour cela, on propose d'introduire une typicité des valeurs. Contrairement à l'intérêt des clusters, la typicité des valeurs n'est pas simplement la contribution d'une valeur à une inertie. En effet, si on étudiait la contribution de chacune des valeurs à l'inertie intra-cluster, on se rendrait vite compte que les valeurs sous-représentées seraient les moins coûteuses à retirer de leur cluster et donc les moins contributrices à l'inertie intra-cluster. On choisit donc de soustraire à cette contribution le coût moyen de réaffectation de la valeur dans chacun des clusters. Ainsi, on ne pénalise pas les valeurs fortement représentées. Intuitivement, une valeur est représentative de son cluster si du point de vue de l'autre variable, le conditionnement par la valeur conduit à une loi similaire au conditionnement par son cluster, et très différente (en moyenne) de celles obtenues en conditionnant par rapport aux autres clusters.

Définition 32. *Pour une valeur $v_{.\lambda}$ appartenant au cluster $c_{.\gamma}$ de la variable-partition X_2^M, la typicité est définie comme l'impact moyen sur l'inertie intra-cluster de la retirer de son cluster et de la réaffecter dans un autre cluster de la partition X_2^M.*

$$\begin{aligned} T_v(v_{.\lambda}) &= \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} P_{X_2^M}(c_{.j})(J_{intra}(X_2^M | c_{.\gamma} \setminus v_{.\lambda}, c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M)) \\ &= \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} P_{X_2^M}(c_{.j})(\xi(\mathcal{M} | c_{.\gamma} \setminus v_{.\lambda}, c_{.j} \cup v_{.\lambda}) - \xi(\mathcal{M})) \end{aligned} \tag{4.30}$$

où $c_{.\gamma} \setminus v_{.\lambda}$ est le cluster $c_{.\gamma}$ duquel on a retiré la valeur $v_{.\lambda}$ et $c_{.j} \cup v_{.\lambda}$ est le cluster $c_{.j}$ dans lequel on a ajouté la valeur $v_{.\lambda}$.

Asymptotiquement, l'inertie intra-cluster d'une partition s'interprète comme une divergence de Jensen-Shannon, qui se décompose en une somme de divergences de Kullback-Leibler sur chacune des valeurs, localement à chaque cluster. On en déduit donc l'estimation asymptotique de l'impact moyen sur

l'inertie intra-cluster de retirer une valeur de son cluster et de la réaffecter dans un autre cluster de la même partition.

Théorème 12. *Asymptotiquement, la typicité d'une valeur $v_{.\lambda}$ appartenant au cluster $c_{.\gamma}$ de la partition X_2^M s'interprète comme une somme de divergences de Kullback-Leibler.*

$$\lim_{m\to+\infty} \frac{T_v(v_{.\lambda})}{m} = -P_{X_2}(v_{.\lambda})KL(P_{X_1^M|v_{.\lambda}}||P_{X_1^M|c_{.\gamma}}) + \frac{1}{1-P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j}\in X_2^M \\ c_{.j}\neq c_{.\gamma}}} P_{X_2^M}(c_{.j})P_{X_2}(v_{.\lambda})KL(P_{X_1^M|v_{.\lambda}}||\alpha_j P_{X_1^M|c_{.j}} + \alpha_\lambda P_{X_1^M|v_{.\lambda}}) \tag{4.31}$$

$$où\ \alpha_\lambda = \frac{P_{X_2}(v_{.\lambda})}{P_{X_2}(v_{.\lambda}) + P_{X_2^M}(c_{.j})}\ et\ \alpha_j = \frac{P_{X_2}(c_{.j})}{P_{X_2}(v_{.\lambda}) + P_{X_2^M}(c_{.j})}$$

Démonstration. La démonstration est détaillée en annexe 4.9.4. □

L'interprétation asymptotique permet de mieux comprendre la typicité d'une valeur : une valeur typique est (i) une valeur dont la fréquence d'apparition dans les données est significative (termes de probabilité d'apparition de la valeur dans les données), (ii) dont la loi conditionnelle associée est très similaire à la loi associée au cluster duquel elle provient (termes de divergence de Kullback-Liebler de la loi de la variable-partition X_1^M conditionnellement à la valeur par rapport à la loi de la variable-partition X_1^M conditionnellement au cluster duquel elle est retirée) et (iii) dont la loi conditionnelle associée est très différente de celle des autres clusters (termes de divergence de Kullback Liebler de la loi de la variable-partition X_1^M conditionnellement à la valeur par rapport à la loi de la variable-partition X_1^M conditionnellement au cluster dans lequel elle est ajoutée). Au contraire, une valeur avec une typicité proche de zéro est soit une valeur tellement peu représentée dans les données qu'elle peut être affectée à n'importe quel cluster sans impact sur la qualité du modèle, soit une valeur significativement représentée dans les données mais qui n'est similaire à aucun cluster.

On a ainsi une mesure qui permet d'évaluer la représentativité d'une valeur au sein de son cluster. La typicité peut être utilisée pour attribuer une étiquette aux clusters. On peut par exemple choisir la valeur la plus typique ou une combinaison de valeurs si plusieurs valeurs ont des typicités très proches.

La mise en place de cet indicateur est assez proche de la méthode *Silhouette* (Rousseeuw, 1987), utilisée pour la validation du clustering. Après avoir appliqué un algorithme basé sur l'optimisation d'une distance, on calcule, pour chaque élément, la distance de l'élément au barycentre de son cluster et la distance moyenne de l'élément à l'ensemble des barycentres des autres clusters. Ces deux termes sont soustraits, ce qui revient à calculer l'impact moyen de la réaffectation d'un élément dans un cluster sur l'inertie intra-cluster du clustering.

La typicité est donc assimilable à Silhouette appliquée à un biclustering et avec pour mesure de dissimilarité la divergence de Kullback-Leibler.

La typicité des valeurs est illustrée sur un biclustering simplifié (voir section 4.3.2) réalisé de la base *Adult*. On se focalise sur la variable *occupation*. Nous avons trois clusters : {*prof-specialty, exec-managerial*}, {*machine-op-inspct, transport-moving, handlers-cleaners, other service, farming-fishing, priv-house-serv, craft-repair*} and {*adm-clerical, protective-serv, sales, tech-support, armed-forces*}. Dans ce cas, la partition est plutôt simple mais les valeurs composant les clusters sont assez différentes. Il est d'ailleurs assez difficile d'attribuer une étiquette à ces trois clusters. En calculant la typicité de chaque valeur, les valeurs les plus représentatives de leur cluster sont respectivement *Prof-specialty, Craft-repair* et *Adm-clerical*. Dans le cas de la base *Adult*, la typicité est très corrélée avec la fréquence d'apparition des valeurs dans les données. Ceci est principalement dû au faible nombre de valeurs dans les clusters : le retrait et l'ajout d'une valeur ont des impacts très importants sur l'inertie intra-cluster car leur poids dans les clusters est significatif. Néanmoins ce résultat est satisfaisant dans le sens où une valeur fréquente joue un rôle important dans les données et donc dans son cluster. La mesure de la typicité montre toute son utilité lorsque les clusters sont plus peuplés et que les fréquences des valeurs sont équilibrées, comme dans le cas d'étude du chapitre 5.

cluster Id	**valeurs**	**typicité**	**Probabilité de la valeur**
Prof-specialty	Prof-specialty	1	18,39%
	Exec-managerial	0,4416	12,46%
Craft-repair	Craft-repair	1	12,51%
	Other-service	0,9016	10,08%
	Machine-op-inspct	0,7146	6,19%
	Transport-moving	0,5174	4,82%
	Handlers-cleaners	0,4842	4,24%
	Farming-fishing	0,2660	3,05%
	Priv-house-serv	0,0684	0,50%
Adm-clerical	Adm-clerical	1	11,49%
	Sales	0,8937	11,27%
	Tech-support	0,3516	2,96%
	Protective-serv	0,1723	2,01%
	Armed-Forces	0,0019	0,03%

TABLE 4.4 – Typicités des valeurs de la variable *occupation* de la base *Adult*, normalisées pour chaque cluster

4.6 Ajout d'une nouvelle valeur

Dans la continuité de l'analyse de la typicité des valeurs, on propose ici de traiter le problème de l'arrivée d'une nouvelle valeur dans les données. L'ajout d'une valeur dans les grilles de co-clustering présente un intérêt particulier dans le cas de l'analyse de base de données de textes. Il est assez courant de traiter ce type de problèmes avec des approches de biclustering où une partition conjointe des textes et des mots est réalisée. Dans le cas où on a construit notre grille de biclustering sur une grande base de textes, on obtient une partition très fine des mots, et asymptotiquement on peut considérer avoir partitionné l'ensemble du lexique de la langue de la base de texte étudiée. Dans ce cas, l'ajout d'une valeur (ici un texte) dans la partition des textes prend du sens. On peut ainsi déployer les résultats sur des nouvelles données, ici des textes.

Le modèle de biclustering $\mathcal{M}$ a été construit avec un ensemble de données $\mathcal{D}$. Considérons que de nouvelles observations soient maintenant à intégrer aux données. Toutes ces observations prennent des valeurs sur X_1 ayant déjà été observées dans $\mathcal{D}$ et une nouvelle et unique valeur sur X_2, que l'on nomme $v_{\cdot\lambda}$. On cherche à intégrer $v_{\cdot\lambda}$ à l'un des clusters de la partition X_2^M.

On a vu que l'inertie totale de la partition X_2^M est constante pour une partition X_1^M fixée. Ainsi, l'ajout d'une valeur au cluster qui détériore le moins l'inertie inter-clusters est également celle qui détériore le moins l'inertie intra-cluster.

Définition 33. *Pour une valeur $v_{\cdot\lambda}$ de la variable X_2, le meilleur cluster d'attribution est celui qui minimise la variation de l'inertie intra-cluster (et donc maximise la variation de l'inertie inter-clusters) lorsque la valeur $v_{\cdot\lambda}$ lui est ajoutée.*

$$\begin{aligned} c_{\cdot j}^* &= \underset{c_{\cdot j} \in X_2^M}{\operatorname{argmax}} (J_{inter}(X_2^M | c_{\cdot j} \cup v_{\cdot\lambda}) - J_{inter}(X_2^M)) \\ &= \underset{c_{\cdot j} \in X_2^M}{\operatorname{argmin}} (J_{intra}(X_2^M | c_{\cdot j} \cup v_{\cdot\lambda}) - J_{intra}(X_2^M)) \\ &= \underset{c_{\cdot j} \in X_2^M}{\operatorname{argmax}} (\xi(\mathcal{M} | c_{\cdot j} \cup v_{\cdot\lambda}) - \xi(\mathcal{M})) \end{aligned} \tag{4.32}$$

Asymptotiquement, l'inertie intra-cluster d'une partition s'interprète comme une divergence de Jensen-Shannon. Contrairement au cas de la typicité, $P_{X_1^M | c_{\cdot j}}$ ne tient pas compte de l'inclusion de la valeur $v_{\cdot\lambda}$ dans le cluster $c_{\cdot j}$. On doit donc prendre en compte la modification de la loi de la variable-partition X_1^M conditionnellement au cluster liée à l'ajout d'une valeur. Le théorème 4.9.1, qui permet de décomposer une divergence de Jensen-Shannon entre plusieurs lois en une somme de divergences de Jensen-Shannon entre deux lois, permet d'obtenir une approximation asymptotique simple de l'impact sur l'inertie intra-cluster de l'ajout d'une nouvelle valeur dans un cluster.

Théorème 13. *Asymptotiquement, la variation d'inertie intra-cluster liée à l'ajout d'une valeur $v_{\cdot\lambda}$ dans un cluster $c_{\cdot j}$ s'interprète comme une divergence*

de Jensen-Shannon.

$$c^*_{.j} \underset{m\to+\infty}{=} \underset{c_{.j}\in X_2^M}{\operatorname{argmin}} \left[P_{X_2^M}(c_{.j})JS(P_{X_1^M|v_{.\lambda}}, P_{X_1^M|c_{.j}}) - P_{X_2}(v_{.\lambda})JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.j}\})\right] \tag{4.33}$$

Démonstration. Le théorème de Huyghens permet de décomposer la divergence de Jensen-Shannon et amène à ce résultat. La démonstration est détaillée en annexe 4.9.5. □

Le premier terme de l'équation 4.33 s'interprète comme la dissimilarité (voir section 4.3.1) entre la valeur et le cluster dans lequel elle a été ajoutée. En minimisant ce terme, on cherche à inclure la valeur dans un cluster similaire. Le second terme est une divergence de Jensen-Shannon entre les lois associée à chaque valeur du cluster, la valeur nouvellement ajoutée étant exclue. On cherche à maximiser ce terme, c'est-à-dire à privilégier un cluster dans lequel les valeurs sont dispersées. C'est là un moyen de préserver les clusters groupant des valeurs très similaires de la dégradation de leur faible inertie intra-cluster.

Nous utilisons la partition simplifiée en trois clusters de la variable *occupation* de la base *Adult* pour illustrer la technique d'ajout d'une valeur dans un cluster. On propose d'insérer trois valeurs différentes dans les clusters de la partition. Le tableau 4.5 présente le coût d'ajout des valeurs dans les clusters. Dans un premier temps, on insère de nouvelles observations prenant toutes la valeur *Other-service/copie* sur X_2. Elles sont réparties sur les clusters de la variable-partition X_1^M de la même manière que les observations prenant la valeur *Other-services*. On essaye d'intégrer la valeur ajoutée *Other-service/copie* dans un des clusters de la partition X_2^M. On observe que l'ajout de cette valeur dans le cluster *Craft-repair* a un coût négatif, ce qui signifie que cet ajout contribue à faire diminuer l'inertie intra-cluster. Ce résultat n'est pas étonnant dans le sens où la valeur *Other-service* a une forte typicité, elle est donc représentative de son cluster. En insérant une valeur exactement identique dans le cluster, on contribue à faire baisser la variance intra-classes, ce qui explique ce résultat.

	Other-service	Uniforme	Marginale
cluster Prof-specialty	$3{,}51.10^{-2}$	$4{,}89.10^{-3}$	$9{,}32.10^{-3}$
cluster Craft-repair	$-7{,}25.10^{-4}$	$3{,}26.10^{-2}$	$1{,}31.10^{-2}$
cluster Adm-clerical	$1{,}08.10^{-2}$	$1{,}46.10^{-2}$	$9{,}54.10^{-4}$

TABLE 4.5 – Coût de l'ajout de la valeur *Other-service*, d'une valeur dont la loi conditionnelle associée est uniforme et d'une valeur dont la loi associée est la même que la loi de la variable-partition X_1^M, dans chacun des clusters de la partition simplifiée de la variable *occupation*.

Dans un second temps, on crée des observations prenant une nouvelle valeur, décrite comme uniforme. On engendre un ensemble de 4 923 observations, c'est-à-dire le nombre d'observations prenant la valeur *Other-service*. Ces

points prennent une seule valeur sur la variable *occupation*. Leurs valeurs sur la variable *education* sont choisies pour que la loi de la variable-partition X_1^M conditionnellement à la valeur ajoutée soit uniforme, ou encore pour que l'entropie de la variable X_1^M conditionnellement à cette nouvelle valeur soit maximale. Dans ce cas, l'ajout de la valeur dans chacun des clusters dégrade l'inertie intra-cluster. Cependant, son ajout dans le cluster *Prof-specialty* minimise cette dégradation. L'entropie de Shannon de la variable-partition X_1^M conditionnellement aux clusters de la variable-partition X_2^M est maximale pour le cluster *Prof-specialty*, ce qui est cohérent avec les hypothèses utilisées pour engendrer les données.

Enfin, on crée une valeur décrite comme marginale. Les 4 923 observations sont engendrées de manière à ce que la loi de X_1^M conditionnellement à la nouvelle valeur soit identique à la loi de la variable-partition $X_1^M : P_{X_1^M|v_{.\lambda}}(c_{i.}) = P_{X_1^M}(c_{i.})$. Le meilleur cluster d'attribution de cette nouvelle valeur est le cluster *Adm-clerical*. Ce cluster est celui qui contribue le plus à l'inertie intra-cluster et le moins à l'inertie inter-clusters. La nouvelle valeur est donc incluse dans un cluster avec une forte variance intra-classe présentant un faible intérêt. Ce type de clusters peut être qualifié de « cluster poubelle ».

4.7 Visualisations

On propose deux visualisations permettant de comprendre comment les clusters intéragissent dans les grilles de co-clustering à deux variables ou plus.

4.7.1 Contribution à l'information mutuelle

On a vu dans le chapitre 2 que l'information mutuelle était une mesure utilisée pour évaluer la qualité d'un biclustering. Cette mesure est d'ailleurs optimisée dans l'approche de Dhillon *et al.* (2003) afin de trouver une grille de biclustering optimale. On a montré également dans la section 4.4.1 que l'inertie inter-clusters des deux partitions d'un biclustering converge asymptotiquement vers l'information mutuelle entre les deux partitions.

L'information mutuelle est une quantité positive. Elle mesure la dépendance entre deux variables.

Définition 34. *L'information mutuelle entre deux variables-partitions X_1^M et X_2^M est définie de la façon suivante (Cover et Thomas, 2006) :*

$$MI(X_1^M; X_2^M) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} MI_{ij} \tag{4.34}$$

avec

$$MI_{ij} = p(c_{ij}) \log \frac{p(c_{ij})}{p(c_{i.})p(c_{.j})} \tag{4.35}$$

Nous proposons une visualisation du biclustering de manière à comprendre comment interagissent les clusters, en exploitant la contribution à l'information mutuelle de chacun des biclusters, c'est-à-dire MI_{ij}. Cette mesure nous permet de comprendre les excès et les déficits d'interactions entre clusters :

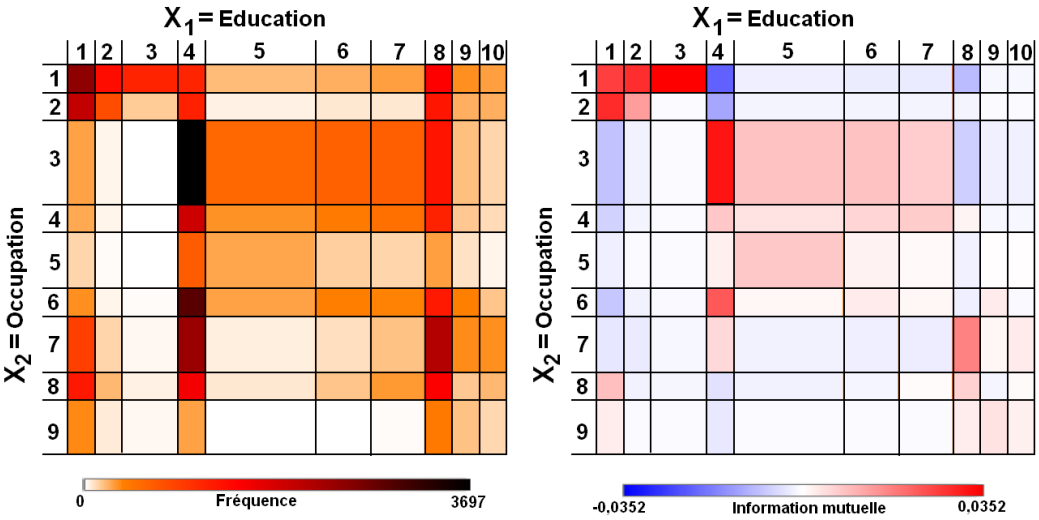
- si $MI_{ij} > 0$: alors cela signifie que $p(c_{ij})$ est supérieure à $p(c_{i.})p(c_{.j})$. On observe un excès d'interactions entre les clusters $c_{i.}$ et $c_{.j}$,
- si $MI_{ij} < 0$: alors cela signifie que $p(c_{ij})$ est inférieure à $p(c_{i.})p(c_{.j})$. On observe un déficit d'interactions entre les clusters $c_{i.}$ et $c_{.j}$,
- si $MI_{ij} = 0$: alors cela signifie soit que $p(c_{ij})$ est égale à $p(c_{i.})p(c_{.j})$, auquel cas soit la quantité d'interactions entre les clusters $c_{i.}$ et $c_{.j}$ est la quantité attendue en cas d'indépendance des partitions ; soit $p(c_{ij}) = 0$, dans ce cas la contribution à l'information mutuelle est nulle car il n'y a pas d'interactions.

Notons que ce type de visualisation a été proposée par Friendly (1994) qui caractérise les excès et déficits d'observations dans un tableau de contingence en étudiant la contribution au test statistique du χ^2. Comme l'information mutuelle, ce test statistique permet de quantifier la corrélation entre deux variables.

Prenons l'exemple de la base Adult pour illustrer l'intérêt de cette visualisation (voir figure 4.9). Les matrices de fréquences et de contributions à l'information mutuelle sont mises en regard afin de montrer les différences d'interprétation qu'on peut avoir avec ces deux représentations. Plus la contribution à l'information mutuelle est élevée plus la cellule est rouge. Plus elle est négative, plus elle est bleue. Lorsque les cellules sont blanches, c'est que la contribution à l'information mutuelle est proche de zéro.

On s'intéresse à deux biclusters en particulier. Le premier est le bicluster $c_{1,3}$ correspondant aux personnes exerçant une profession spécialisée et ayant un niveau d'étude de type doctorat ou formation spécialisée. L'autre est le bicluster $c_{1,4}$ correspondant aux personnes exerçant une profession spécialisée et ayant arrêté les études au lycée. On observe dans ces deux biclusters des effectifs très proches (1 156 pour $c_{1,3}$ et 1 159 pour $c_{1,4}$), ce qui explique leur couleur commune dans la matrice des fréquences. Cependant, la matrice de contribution à l'information mutuelle nous montre qu'on a un excès d'observations dans le bicluster $c_{1,3}$ (il est coloré en rouge) et un déficit dans $c_{1,4}$ (il est coloré en bleu). Pour mieux comprendre ce résultat, détaillons les calculs :

Les excès. 18,39% des personnes de la base Adult exercent des professions spécialisés et 2,92% ont un niveau d'étude de type doctorat ou formation spécialisée. On s'attend donc, en cas d'indépendance des partitions, à observer $p(c_{1.}) \times p(c_{.3}) = 0{,}54\%$ de personnes avec ce niveau de formation occupant ce type d'emplois, c'est-à-dire environ 263 personnes. Or on en observe 1 159, soit 4,4 fois plus qu'attendu, ce qui donne une valeur de la contribution à l'information mutuelle positive de 0,0352. On a donc un excès important de



(a) Fréquence (b) Contribution à l'information mutuelle

FIGURE 4.9 – Fréquence et contribution à l'information mutuelle des biclusters formés par les partitions des variables occupation et education.

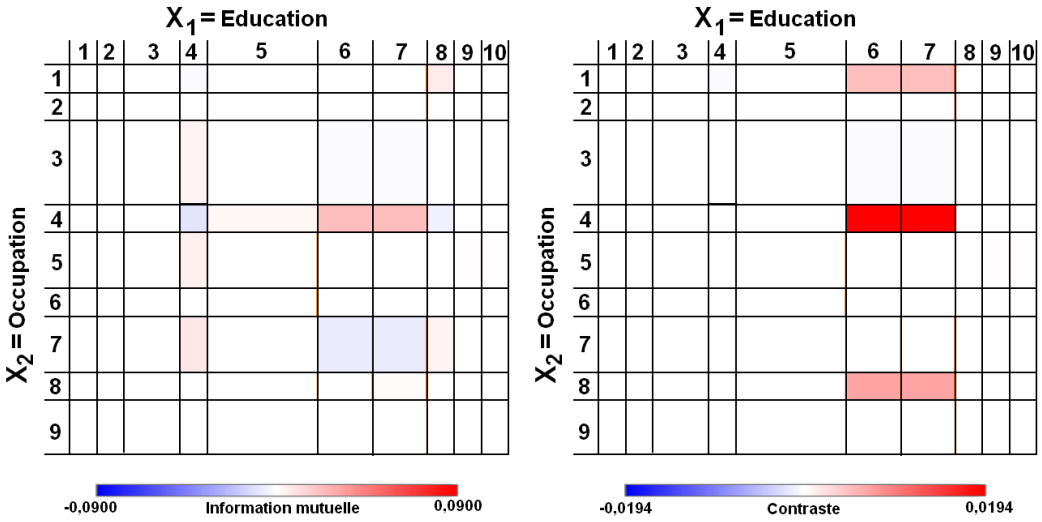
personnes dans cette catégorie professionnelle avec ce niveau d'instruction.

Les déficits. 18,39% des personnes de la base Adult exercent des professions spécialisés et 32,32% ont un niveau de formation équivalent au lycée. On s'attend donc, en cas d'indépendance des partitions, à observer $p(c_{1.}) \times p(c_{.4}) = 5{,}94\%$ de personnes avec ce niveau de formation occupant ce type d'emplois, c'est-à-dire environ 2 902 personnes. Or on en observe 1 159, soit 2,5 fois moins qu'attendu, ce qui donne une valeur de la contribution à l'information mutuelle négative de $-0{,}0218$. On a donc un déficit important de personnes dans cette catégorie professionnelle avec ce niveau d'instruction.

La contribution à l'information mutuelle nous permet donc de comprendre entre quels clusters on observe des excès ou déficits de co-occurrences par rapport à la quantité attendue. Cette mesure présente, d'autre part, l'avantage de tenir compte de l'effectif du bicluster. Ainsi, un bicluster très peu peuplé ne contribue pas de manière significative à l'information mutuelle, même si la fréquence observée est très différente de la fréquence attendue.

4.7.2 Fonction de contraste

Le contraste est une mesure dérivée de l'information mutuelle permettant de visualiser les excès et déficits d'interactions entre clusters lorsque la grille de co-clustering est définie par plus de deux partitions. Dans le cas du biclustering, la visualisation matricielle est adaptée. Dans le cas où on a plus de deux



(a) Contribution à l'information mutuelle (b) Contraste

FIGURE 4.10 – Contribution à l'information mutuelle et contraste des biclusters formés par les partitions des variables occupation et éducation.

dimensions, visualiser les interactions entre clusters peut vite s'avérer complexe. On choisit donc de conserver une visualisation matricielle en deux dimensions et on introduit une mesure de contraste. Pour cela, nous choisissons deux partitions à visualiser dans la matrice, et nous fixons les clusters des autres partitions que nous appelons *contexte*. Le contraste nous permet donc de visualiser les excès et les déficits d'observations entre un segment des données fixé (le contexte) et les biclusters affiché dans la matrice.

Définition 35. *Le contraste entre le couple de variables-partitions à visualiser (X_1^M et X_2^M) et le contexte $c_{..i_3...i_d}$ est défini de la manière suivante :*

$$MI[(X_1^M, X_2^M); c_{..i_3...i_d}] = \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} MI_{i_1 i_2; i_3...i_d} \tag{4.36}$$

avec

$$MI_{i_1 i_2; i_3...i_d} = p(c_{i_1 i_2...i_d}) \log \frac{p(c_{i_1 i_2...i_d})}{p(c_{i_1 i_2...}) p(c_{..i_3...i_d})} \tag{4.37}$$

Prenons l'exemple de la base *Adult* pour illustrer l'intérêt de cette visualisation (voir figure 4.10). La matrice de contribution à l'information mutuelle entre les partitions des variables *occupation* et *education* conditionnellement à l'*age*, et la matrice du contraste sont mises en regard afin de montrer les différences d'interprétation qu'on peut avoir avec ces deux représentations. Dans le cas présent on a fixé le contexte (ici l'age) sur l'intervalle 17-19 ans.

Les deux matrices n'apportent pas la même explication des relations entre les clusters. Intéressons-nous au cluster $c_{8,6,.}$. Il s'agit des personnes occupant des postes dans le commerce ayant un niveau d'étude correspondant au collège. Dans le cas de l'information mutuelle conditionnelle, on n'observe ni excès ni déficit de co-occurences alors que dans le cas du contraste, on observe un excès relativement important.

Information mutuelle conditionnelle. Parmi les personnes de 17 à 19 ans, 66,34% d'entre elles ont un niveau d'étude correspondant au collège et 21,95% occupent un poste dans le commerce. On s'attend donc, en cas d'indépendance des partitions, à observer $p(c_{8..}|c_{..i_3} = 17-19) \times p(c_{.6.}|c_{..i_3} = 17-19) = 14{,}56\%$ des personnes de cet age avec ce niveau de formation occupant ce type d'emplois, c'est-à-dire environ 212 personnes et on en observe 205. Le déficit est donc très léger.

Contraste. 2,98% des personnes de la base Adult ont entre 17 et 19 ans. 3,08% des individus de la base ont un niveau d'étude correspondant au collège et occupent un poste dans le commerce. On s'attend donc, en cas d'indépendance des partitions, à observer $p(c_{..i_3} = 17-19) \times p(c_{8..}, c_{.6.}) = 0{,}09\%$ de personnes de cet age avec ce niveau de formation occupant ce type d'emplois, c'est-à-dire environ 45 personnes et on en observe 205. On a donc un important excès : 4,5 fois la fréquence attendue.

Dans le cas de l'information mutuelle, si on se focalise sur le segment de la population âgée de 17 à 19 ans, on n'observe ni excès ni déficit d'individus ayant un niveau de formation équivalent au collège et exerçant un emploi dans le commerce. Cependant, si on s'intéresse à toute la population de la base Adult, on observe une très forte corrélation entre la tranche d'age $17-19$ ans et les personnes occupant un poste dans la vente avec un niveau de formation correspondant au collège, qui se traduit par un excès d'observations. Il n'y a pas une interprétation meilleure que l'autre. On propose simplement une analyse différente des excès et des déficits d'interactions entre les clusters dans le cas où on a plus de deux dimensions.

4.8 Bilan et généralisation au cas du co-clustering à d dimensions

Cette section récapitule les différentes mesures introduites dans le chapitre et propose de les étendre au cas général, c'est-à-dire au cas où on a d variables qui décrivent les données. On a vu dans ce chapitre différents aspects de l'analyse exploratoire d'un biclustering. Dans un premier temps, nous avons fait émerger du critère MODL une mesure de similarité qui s'interprète comme une divergence de Jensen-Shannon entre les lois conditionnelle associées aux clusters de la partition, pondérée par le poids clusters. Nous avons utilisé cette

mesure de dissimilarité afin de simplifier le biclustering via une classification hiérarchique ascendante. Nous avons ensuite défini les notions d'intérêt de typicité. Dans le cas où nous avons une partition fine des variables, nous allons étudier l'intérêt des clusters afin de détecter les clusters atypiques. Ce type d'études est mise en œuvre lorsqu'on s'intéresse aux phénomènes de niches dans les données. Lorsque les clusters sont peuplés, nous proposons une mesure de typicité permettant de déterminer la valeur la plus représentative de son cluster. Cette valeur peut servir d'étiquette au cluster. Enfin des visualisations conduisant à différentes interprétations sont proposées. Toutes ces mesures ont été déduites du critère MODL et leurs estimations asymptotiques s'interprètent le plus souvent comme des divergences de probabilités. Le tableau des pages 101 et 102 est un récapitulatif de toutes ces mesures, de leurs définitions, de leurs formulations exactes et de leurs propriétés asymptotiques.

La définition du critère MODL dans le cas général (voir section 2.2.5) est adaptée à des données décrites par plusieurs variables pouvant être soit continues, soit nominales. Les différentes mesures introduites dans ce chapitre sont directement déduites du critère optimisé dans l'approche MODL et donc peuvent également s'adapter dans le cas général avec plusieurs dimensions. La formulation de tous les indicateurs est inchangée et ne dépend pas du nombre de dimensions.

Dans le cas de l'étude asymptotique d'une grille de biclustering, les indicateurs s'interprètent comme des divergences entre les lois de probabilité d'une variable-partition conditionnellement à un cluster $P_{X_1^M|c.i}$ ou conditionnellement à une valeur $P_{X_1^M|v.i}$. Si on généralise au co-clustering en d dimensions, les indicateurs ont la même interprétation asymptotique qu'avec deux dimensions : des divergences entre les lois de probabilités jointes des $d-1$ variables-partitions non étudiées, conditionnellement à un cluster de la partition étudiée $P_{X_1^M,X_2^M,...,X_{d-1}|c_{...i}}$ ou conditionnellement à une valeur de la variable étudiée $P_{X_1^M,X_2^M,...,X_{d-1}|v_{...i}}$. Ainsi, il est aisé d'adapter les indicateurs introduits dans ce chapitre au cas du co-clustering avec d variables. Notons que l'égalité des inerties inter-clusters ($J_{inter}(X_1^M) = J_{inter}(X_2^M)$) ne se vérifie plus lorsqu'il y a plus de deux dimensions. En effet, asymptotiquement et dans le cas général, l'inertie inter-clusters de la variable-partition X_d^M s'interprète comme l'information mutuelle entre la partition étudiée et l'intersection de toutes les autres partitions : $MI(X_d^M, X_1^M \times X_2^M \times ... \times X_{d-1}^M)$.

La définition de certains des indicateurs n'a pas de sens dans le cas où la variable est continue. Ainsi, on n'étudie pas l'inertie intra-cluster, la typicité des valeurs et l'ajout d'une valeur dans le cas de variables continues.

L'utilisation de ces outils d'analyse exploratoire est illustrée dans le chapitre 5. Il s'agit de plusieurs analyses de données volumineuses, pour lesquelles il est nécessaire de disposer des indicateurs introduits dans ce chapitre afin d'interpréter les résultats. Nous proposons également une méthodologie d'analyse dans le chapitre 5 afin de définir un contexte pour l'utilisation de chacun des outils d'analyse exploratoire.

Mesure	**Description**	**Formalisation**	**Comportement asymptotique**
Entropie de Shannon	Mesure l'information présente dans une loi tirée suivant une variable aléatoire X_1.	$H(X_1) = -\sum_{i=1}^{n_1} P_{X_1}(v_{i.}) \log P_{X_1}(v_{i.})$	
Divergence de Kullback-Leibler	Mesure l'information perdue lorsque une loi P_2 est utilisée pour estimer une loi P_1	$KL(P_{X_1^M\|c_{.1}} \| P_{X_1^M\|c_{.2}}) = \sum_{i=1}^{k_1} P_{X_1^M\|c_{.1}}(c_{i.}) \log \frac{P_{X_1^M\|c_{.1}}(c_{i.})}{P_{X_1^M\|c_{.2}}(c_{i.})}$	
Divergence de Jensen-Shannon	Somme pondérée de divergences de Kullback-Leibler symétrisées	$JS(P_{X_1^M\|c_{.1}}, ..., P_{X_1^M\|c_{.k_2}}) = \sum_{i=1}^{k_2} \alpha_{.i} KL(P_{X_1^M\|c_{.i}} \| \alpha_{.1} P_{X_1^M\|c_{.1}} + ... + \alpha_{.k_2} P_{X_1^M\|c_{.k_2}})$	
Dissimilarité entre clusters	Coût lié à la fusion de deux clusters	$\Delta(c_{.1}, c_{.2}) = \xi(\mathcal{M}_{(c_{.1} \cup c_{.2})}) - \xi(\mathcal{M}^*)$	$\lim_{m \to +\infty} \frac{\Delta(c_{1.}, c_{2.})}{m} = (P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2})) JS(P_{X_1^M\|c_{.1}}, P_{X_1^M\|c_{.2}})$
Inertie inter-clusters	Mesure la dispersion moyenne des clusters	$J_{inter}(X_2^M) = \xi(\mathcal{M}_{X_1^M \emptyset}) - \xi(\mathcal{M}^*)$	$\lim_{m \to +\infty} \frac{J_{inter}(X_2^M)}{m} = JS(P_{X_1^M\|c_{.1}}, ..., P_{X_1^M\|c_{.k_2}})$ $\lim_{m \to +\infty} \frac{J_{inter}(X_2^M)}{m} \in [0, H(X_2^M)]$ $\lim_{m \to +\infty} \frac{J_{inter}(X_2^M)}{m} = I(X_1^M; X_2^M)$

Inertie intra-cluster	Mesure la dispersion moyenne des valeurs dans les clusters	$J_{intra}(X_2^M) = \xi(\mathcal{M}^*) - \xi(\mathcal{M}_{X_1^M \infty})$	$\lim_{m\to+\infty} \frac{J_{intra}(X_2^M)}{m} = \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) \underset{\forall v_{.i} \in c_{.j}}{JS} (P_{X_1^M \mid v_{.i}})$ $\lim_{m\to+\infty} \frac{J_{intra}(X_2^M)}{m} \in \left[0, \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j}) H(X_2 \mid c_{.j})\right]$ $\lim_{m\to+\infty} \frac{J_{inter}(X_2^M)}{m} = \sum_{j=1}^{k_2} I(X_1^M; X_2 \mid c_{.j})$
Inertie totale	Somme des inerties inter- et intra-cluster, mesure indépendante de la partition X_2^M qu'elle caractérise	$J(X_2^M) = J_{inter}(X_2^M) + J_{intra}(X_2^M)$	$\lim_{m\to+\infty} \frac{J(X_2^M)}{m} = JS(P_{X_1^M \mid v_{.1}}, P_{X_1^M \mid v_{.2}}, ..., P_{X_1^M \mid v_{.n_2}})$
Intérêt du cluster	Mesure l'éloignement du cluster de la moyenne des données	$T_c(c_{.\gamma}) = J_{inter}(X_2^M) - J_{inter}(X_2^M \setminus c_{.\gamma})$	$\lim_{m\to+\infty} \frac{T_c(c_{.\gamma})}{m} = P_{X_2^M}(c_{.\gamma}) KL(P_{X_1^M \mid c_{.\gamma}} \mid\mid P_{X_1^M})$
Typicité d'une valeur	Mesure la représentativité d'une valeur au sein de son cluster	$T_v(v_{.\lambda}) = -J_{intra}(X_2^M)$ $+ \sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} \frac{P_{X_2^M}(c_{.j})}{1 - P_{X_2^M}(c_{.\gamma})} J_{intra}(X_2^M \mid c_{.\gamma} \setminus v_{.\lambda}, c_{.j} \cup v_{.\lambda})$	$\lim_{m\to+\infty} \frac{T_v(v_{.\lambda})}{m} = -P_{X_2}(v_{.\lambda}) \Big(KL(P_{X_1^M \mid v_{.\lambda}} \mid\mid P_{X_1^M \mid c_{.\gamma}}) +$ $\sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} \frac{P_{X_2^M}(c_{.j})}{1 - P_{X_2^M}(c_{.\gamma})} KL(P_{X_1^M \mid v_{.\lambda}} \mid\mid \alpha_j P_{X_1^M \mid c_{.j}} + \alpha_\lambda P_{X_1^M \mid v_{.\lambda}}) \Big)$
Coût d'ajout	Mesure l'impact de l'ajout d'une valeur extérieure au modèle dans un cluster	$c_{.j}^* = \underset{c_{.j} \in X_2^M}{\text{argmax}} (J_{inter}(X_2^M \mid c_{.j} \cup v_{.\lambda}) - J_{inter}(X_2^M))$ $= \underset{c_{.j} \in X_2^M}{\text{argmin}} (J_{intra}(X_2^M \mid c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M))$	$c_{.j}^* = \underset{\substack{c_{.j} \in X_2^M \\ m\to+\infty}}{\text{argmin}} \big[P_{X_2^M}(c_{.j}) JS(P_{X_1^M \mid v_{.\lambda}}, P_{X_1^M \mid c_{.j}})$ $- P_{X_2}(v_{.\lambda}) JS(\{P_{X_1^M \mid v_{.i}}, \forall v_{.i} \in c_{.j}\}) \big]$

4.9 Annexes

4.9.1 Décomposition de la divergence de Jensen-Shannon

La divergence de Jensen-Shannon généralisée de k lois peut se décomposer en une somme de divergences de Jensen-Shannon généralisées de deux lois :

$$JS^{\alpha_1,\alpha_2,...,\alpha_k}(P_1,P_2,...,P_k) = (1-\alpha_k)JS^{\frac{\alpha_1}{1-\alpha_k},\frac{\alpha_2}{1-\alpha_k},...,\frac{\alpha_{k-1}}{1-\alpha_k}}(P_1,P_2,...P_{k-1}) + JS^{1-\alpha_k,\alpha_k}\left(\sum_{j=1}^{k-1}\frac{\alpha_j}{1-\alpha_k}P_j, P_k\right)$$

avec α_i le coefficient de mélange associé à la loi P_i, tel que $\sum_{i=1}^{k}\alpha_i = 1$.

Démonstration. Soient trois lois P_1, P_2 et P_3 auxquelles sont associés les coefficients de mélange α_1, α_2 et α_3 tels que $\alpha_1+\alpha_2+\alpha_3 = 1$. (Rappel : $JS^{\alpha_1,\alpha_2}(P_1,P_2) = H(\alpha_1P_1+\alpha_2P_2) - \alpha_1H(P_1) - \alpha_2H(P_2)$.

$$\begin{aligned}
&JS^{1-\alpha_3,\alpha_3}\left(\frac{\alpha_1}{1-\alpha_3}P_1+\frac{\alpha_2}{1-\alpha_3}P_2, P_3\right) + (1-\alpha_3)JS^{\frac{\alpha_1}{1-\alpha_3},\frac{\alpha_2}{1-\alpha_3}}(P_1,P_2)\\
&= -(\alpha_1P_1+\alpha_2P_2+\alpha_3P_3)\log(\alpha_1P_1+\alpha_2P_2+\alpha_3P_3)\\
&\quad + (\alpha_1P_1+\alpha_2P_2)\log\left(\frac{1}{1-\alpha_3}(\alpha_1P_1+\alpha_2P_2)\right) + \alpha_3P_3\log(P_3)\\
&- (\alpha_1P_1+\alpha_2P_2)\log\left(\frac{1}{1-\alpha_3}(\alpha_1P_1+\alpha_2P_2)\right)\\
&\quad + \alpha_1P_1\log(P_1) + \alpha_2P_1\log(P_2)\\
&= \alpha_1P_1(\log(P_1) - \log(\alpha_1P_1+\alpha_2P_2+\alpha_3P_3))\\
&\quad + \alpha_2P_2(\log(P_2) - \log(\alpha_1P_1+\alpha_2P_2+\alpha_3P_3))\\
&\quad + \alpha_3P_3(\log(P_3) - \log(\alpha_1P_1+\alpha_2P_2+\alpha_3P_3))\\
&= JS^{\alpha_1,\alpha_2,\alpha_3}(P_1,P_2,P_3)
\end{aligned}$$

Soient k lois $P_1, P_2, ..., P_k$ avec les coefficients de mélanges associés $\beta_1, \beta_2, ..., \beta_k$ tels que $\sum_{i=1}^{k}\beta_i = 1$. Soient $\alpha_1, \alpha_2, ..., \alpha_{k-1}$ tels que $\alpha_i = \frac{\beta_i}{1-\beta_k}$ et $\sum_{i=1}^{k-1}\alpha_i = 1$ les coefficients d'un mélange des $k-1$ premières lois. On fait l'hypothèse que le théorème est vrai pour un mélange de $k-1$ lois.

$$
\begin{aligned}
&JS^{1-\beta_k,\beta_k}\left(\frac{\beta_1}{1-\beta_k}P_1+\frac{\beta_2}{1-\beta_k}P_2+...+\frac{\beta_{k-1}}{1-\beta_k}P_{k-1},P_k\right)\\
&\quad+(1-\beta_k)JS^{\frac{\beta_1}{1-\beta_k},\frac{\beta_2}{1-\beta_k},...,\frac{\beta_{k-1}}{1-\beta_k}}(P_1,P_2,...,P_{k-1})\\
&=JS^{1-\beta_k,\beta_k}\left(\frac{\beta_1}{1-\beta_k}P_1+\frac{\beta_2}{1-\beta_k}P_2+...+\frac{\beta_{k-1}}{1-\beta_k}P_{k-1},P_k\right)\\
&\quad+(1-\beta_k)\left[JS^{1-\alpha_{k-1},\alpha_{k-1}}\left(\frac{\alpha_1}{1-\alpha_{k-1}}P_1+\frac{\alpha_2}{1-\alpha_{k-1}}P_2+...+\frac{\alpha_{k-2}}{1-\alpha_{k-1}}P_{k-2},P_{k-1}\right)\right.\\
&\quad+...\\
&\quad+(\alpha_1+\alpha_2+\alpha_3)JS^{\frac{\alpha_1+\alpha_2}{\alpha_1+\alpha_2+\alpha_3},\frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3}}\left(\frac{\alpha_1}{\alpha_1+\alpha_2}P_1+\frac{\alpha_2}{\alpha_1+\alpha_2}P_2,P_3\right)\\
&\quad\left.+(\alpha_1+\alpha_2)JS^{\frac{\alpha_1}{\alpha_1+\alpha_2},\frac{\alpha_2}{\alpha_1+\alpha_2}}(P_1,P_2)\right]\\
&=JS^{\beta_1+\beta_2+...+\beta_{k-1},\beta_k}\left(\frac{\beta_1}{\beta_1+\beta_2+...+\beta_{k-1}}P_1\right.\\
&\quad\left.+\frac{\beta_2}{\beta_1+\beta_2+...+\beta_{k-1}}P_2+...+\frac{\beta_{k-1}}{\beta_1+\beta_2+...+\beta_{k-1}}P_{k-1},P_k\right)\\
&\quad+(\beta_1+\beta_2+...+\beta_{k-1})\left[JS^{\frac{\beta_1+\beta_2+...+\beta_{k-2}}{\beta_1+\beta_2+...+\beta_{k-1}},\frac{\beta_{k-1}}{\beta_1+\beta_2+...+\beta_{k-1}}}\left(\frac{\beta_1}{\beta_1+\beta_2+...+\beta_{k-2}}P_1\right.\right.\\
&\quad\left.+\frac{\beta_2}{\beta_1+\beta_2+...+\beta_{k-2}}P_2+...+\frac{\beta_{k-2}}{\beta_1+\beta_2+...+\beta_{k-2}}P_{k-2},P_{k-1}\right)\\
&\quad+...\\
&\quad\left.+\frac{\beta_1+\beta_2+\beta_3}{\beta_1+\beta_2+...+\beta_{k-1}}JS^{\frac{\beta_1+\beta_2}{\beta_1+\beta_2+\beta_3},\frac{\beta_3}{\beta_1+\beta_2+\beta_3}}\left(\frac{\beta_1}{\beta_1+\beta_2}P_1+\frac{\beta_2}{\beta_1+\beta_2}P_2,P_3\right)\right.\\
&+\left.\frac{\beta_1+\beta_2}{\beta_1+\beta_2+...+\beta_{k-1}}JS^{\frac{\beta_1}{\beta_1+\beta_2},\frac{\beta_2}{\beta_1+\beta_2}}(P_1,P_2)\right]\\
&=JS^{\beta_1+\beta_2+...+\beta_{k-1},\beta_k}\left(\frac{\beta_1}{\beta_1+\beta_2+...+\beta_{k-1}}P_1+\frac{\beta_2}{\beta_1+\beta_2+...+\beta_{k-1}}P_2\right.\\
&\quad\left.+...+\frac{\beta_{k-1}}{\beta_1+\beta_2+...+\beta_{k-1}}P_{k-1},P_k\right)\\
&\quad+(\beta_1+\beta_2+...+\beta_{k-1})JS^{\frac{\beta_1+\beta_2+...+\beta_{k-2}}{\beta_1+\beta_2+...+\beta_{k-1}},\frac{\beta_{k-1}}{\beta_1+\beta_2+...+\beta_{k-1}}}\left(\frac{\beta_1}{\beta_1+\beta_2+...+\beta_{k-2}}P_1\right.\\
&\quad\left.+\frac{\beta_2}{\beta_1+\beta_2+...+\beta_{k-2}}P_2+...+\frac{\beta_{k-1}}{\beta_1+\beta_2+...+\beta_{k-2}}P_{k-2},P_{k-1}\right)\\
&\quad+...\\
&\quad+(\beta_1+\beta_2+\beta_3)JS^{\frac{\beta_1+\beta_2}{\beta_1+\beta_2+\beta_3},\frac{\beta_3}{\beta_1+\beta_2+\beta_3}}\left(\frac{\beta_1}{\beta_1+\beta_2}P_1+\frac{\beta_1}{\beta_1+\beta_2}P_2,P_3\right)\\
&\quad+(\beta_1+\beta_2)JS^{\frac{\beta_1}{\beta_1+\beta_2},\frac{\beta_2}{\beta_1+\beta_2}}(P_1,P_2)\\
&=JS^{\beta_1,\beta_2,...,\beta_k}(P_1,P_2,...,P_k)
\end{aligned}
$$

Par récurrence, le théorème est démontré.

□

4.9.2 Interprétation asymptotique du coût de fusion de deux clusters comme une divergence de Jensen-Shannon

Soient $c_{.1}$ et $c_{.2}$ deux clusters issus de la variable-partition X_2^M de la variable X_2. Dans le régime asymptotique, la dissimilarité $\Delta(c_{.1}, c_{.2})$ entre ces deux clusters s'interprète comme la divergence de Jensen-Shannon entre les lois $P_{X_1^M|c_{.1}}$ et $P_{X_1^M|c_{.2}}$:

$$\lim_{m\to+\infty} \frac{\Delta(c_{1.}, c_{2.})}{m} = (P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2}))JS^{\alpha_1,\alpha_2}(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}})$$

où $\alpha_1 = \dfrac{P_{X_2^M}(c_{.1})}{P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2})}$ et $\alpha_2 = \dfrac{P_{X_2^M}(c_{.2})}{P_{X_2^M}(c_{.1}) + P_{X_2^M}(c_{.2})}$

Démonstration. La variation du critère lié à la fusion des clusters $c_{.1}$ et $c_{.2}$ issus de la partition X_1^M de la variable X^1 est définie de la manière suivante :

$$\Delta(c_{.1}, c_{.2}) = \xi(\mathcal{M}_{(c_{.1}\cup c_{.2})}) - \xi(\mathcal{M}_{c_{.1},c_{.2}})$$

Asymptotiquement, l'impact sur l'a priori peut être négligé, on ne s'intéresse donc qu'à la variation de la vraisemblance :

$$\begin{aligned}\mathcal{L}(\mathcal{M}) =& \log(m!) - \sum_{\substack{c_i\in X_1^M\\ c_j\in X_2^M}} \log(m_{ij}^c!) + \sum_{c_j\in X_2^M} \log(m_{.j}^c!) - \sum_{x_j\in X_2} \log(m_{.j}!)\\ &+ \sum_{c_i\in X_1^M} \log(m_{i.}^c!) - \sum_{x_i\in X_1} \log(m_{i.}!)\end{aligned}$$

pour $m \to \infty$.

Seuls les termes se référant à la partition X_1^M sont impactés :

$$\begin{aligned}\Delta(c_{.1}, c_{.2}) =& - \sum_{\substack{c_i\in X_1^M\\ c_j\in X_2^M\\ c_i\neq\{c_{.1},c_{.2}\}}} \log(m_{ij}^c!) - \sum_{c_j\in X_2^M} \log((m_{1j}^c + m_{2j}^c)!)\\ &+ \sum_{\substack{c_i\in X_1^M\\ c_j\in X_2^M\\ c_i\neq\{c_{.1},c_{.2}\}}} \log(m_{ij}^c!) - \sum_{c_j\in X_2^M} (\log(m_{1j}^c!) + \log(m_{2j}^c!))\\ &+ \sum_{\substack{c_i\in X_1^M\\ c_i\neq\{c_{.1},c_{.2}\}}} \log(m_{i.}^c!) + \log((m_{1.}^c + m_{2.}^c)!)\\ &- \sum_{\substack{c_i\in X_1^M\\ c_i\neq\{c_{.1},c_{.2}\}}} \log(m_{i.}^c!) - \log(m_{1.}^c!) - \log(m_{2.}^c!)\end{aligned}$$

Asymptotiquement, nous appliquons l'approximation de Stirling : $\log(m!) = m\log(m)$

$$\begin{aligned}
\Delta(c_{.1}, c_{.2}) &= -\sum_{c_j \in X_2^M} \log((m^c_{1j} + m^c_{2j})!) - \sum_{c_j \in X_2^M} (\log(m^c_{1j}!) + \log(m^c_{2j}!)) \\
&\quad + \log((m^c_{1.} + m^c_{2.})!) - \log(m^c_{1.}!) - \log(m^c_{2.}!) \\
\Delta(c_{.1}, c_{.2}) &= -\sum_{c_j \in X_2^M} (m^c_{1j} + m^c_{2j}) \log(m^c_{1j} + m^c_{2j}) - \sum_{c_j \in X_2^M} (m^c_{1j} \log(m^c_{1j}) + m^c_{2j} \log(m^c_{2j})) \\
&\quad + (m^c_{1.} + m^c_{2.}) \log(m^c_{1.} + m^c_{2.}) - m^c_{1.} \log(m^c_{1.}) - m^c_{2.} \log(m^c_{2.}) \\
\Delta(c_{.1}, c_{.2}) &= -\sum_{c_j \in X_2^M} (m^c_{1j} + m^c_{2j}) \log((m^c_{1j} + m^c_{2j})) - \sum_{c_j \in X_2^M} (m^c_{1j} \log(m^c_{1j}) + m^c_{2j} \log(m^c_{2j})) \\
&\quad + \sum_{c_j \in X_2^M} (m^c_{1j} + m^c_{2j}) \log((m^c_{1.} + m^c_{2.})) - \sum_{c_j \in X_2^M} m^c_{1j} \log(m^c_{1.}) - \sum_{c_j \in X_2^M} m^c_{2j} \log(m^c_{2.}) \\
\Delta(c_{.1}, c_{.2}) &= -\sum_{c_j \in X_2^M} (m^c_{1j} + m^c_{2j}) \log \frac{m^c_{1j} + m^c_{2j}}{m^c_{1.} + m^c_{2.}} - \sum_{c_j \in X_2^M} m^c_{1j} \log \frac{m^c_{1j}}{m^c_{1.}} + \sum_{c_j \in X_2^M} m^c_{2j} \log \frac{m^c_{2j}}{m^c_{2.}}
\end{aligned}$$

Soient $P_{X_1^M|c_{.1}}$ et $P_{X_1^M|c_{.2}}$ les lois de X_1^M conditionnellement à $c_{.1}$ et $c_{.2}$ telles que :

$$P_{X_1^M|c_{.1}}(j) = \frac{m^c_{1j}}{m^c_{1.}} \text{ et } P_{X_1^M|c_{.2}}(j) = \frac{m^c_{2j}}{m^c_{2.}}$$

Soient α_1 et α_2 des coefficients de mélange des lois $P_{X_1^M|c_{.1}}$ et $P_{X_1^M|c_{.2}}$:

$$\alpha_1 = \frac{m^c_{1.}}{m^c_{1.} + m^c_{2.}} \text{ et } \alpha_2 = \frac{m^c_{2.}}{m^c_{1.} + m^c_{2.}}$$

On obtient alors :

$$\begin{aligned}
\Delta(c_{.1}, c_{.2}) &= (m^c_{1.} + m^c_{2.}) H(\alpha_1 P_{X_1^M|c_{.1}} + \alpha_2 P_{X_1^M|c_{.2}}) - m^c_{1.} H(P_{X_1^M|c_{.1}}) - m^c_{2.} H(P_{X_1^M|c_{.2}}) \\
\Delta(c_{.1}, c_{.2}) &= (m^c_{1.} + m^c_{2.}) \Big(H(\alpha_1 P_{X_1^M|c_{.1}} + \alpha_2 P_{X_1^M|c_{.2}}) \\
&\quad - \frac{m^c_{1.}}{m^c_{1.} + m^c_{2.}} H(P_{X_1^M|c_{.1}}) - \frac{m^c_{2.}}{m^c_{1.} + m^c_{2.}} H(P_{X_1^M|c_{.2}}) \\
\Delta(c_{.1}, c_{.2}) &= (m^c_{1.} + m^c_{2.}) \Big(H(\alpha_1 P_{X_1^M|c_{.1}} + \alpha_2 P_{X_1^M|c_{.2}}) - \alpha_1 H(P_{X_1^M|c_{.1}}) - \alpha_2 H(P_{X_1^M|c_{.2}}) \Big) \\
\Delta(c_{.1}, c_{.2}) &= (m^c_{1.} + m^c_{2.}) JS^{\alpha_1,\alpha_2}(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}}) \\
\frac{\Delta(c_{.1}, c_{.2})}{m} &= (\alpha_1 + \alpha_2) JS^{\alpha_1,\alpha_2}(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}})
\end{aligned}$$

□

4.9.3 Interprétation asymptotique de l'inertie inter-clusters comme une information mutuelle

L'inertie inter-cluster J_{inter} est asymptotiquement proportionnelle à l'information mutuelle entre les variables-partitions X_1^M et X_2^M :

$$\lim_{m\to+\infty} \frac{J_{inter}\left(X_2^M\right)}{m} = MI\left(X_1^M; X_2^M\right)$$

Démonstration. Asymptotiquement, l'inertie inter-clusters tend vers la divergence de Jensen-Shannon entre les lois de X_1^M conditionnellement aux clusters de la partition X_2^M :

$$\begin{aligned}\lim_{m\to+\infty} \frac{J_{inter}\left(X_2^M\right)}{m} &= JS^{\alpha_1,\alpha_2,...,\alpha_{k_2}}\left(P_{X_1^M|c_{.1}}, P_{X_1^M|c_{.2}}, ..., P_{X_1^M|c_{.k_2}}\right) \\ &= H\left(\sum_{j=1}^{k_2} \alpha_j P_{X_1^M|c_{.j}}\right) - \sum_{j=1}^{k_2} \alpha_j H\left(P_{X_1^M|c_{.j}}\right)\end{aligned}$$

avec $\alpha_j = P_{X_2^M}(c_{.j})$.

On a alors :

$$\begin{aligned}H\left(\sum_{j=1}^{k_2} \alpha_j P_{X_1^M|c_{.j}}\right) &= H\left(\sum_{j=1}^{k_2} P_{X_2^M}(c_{.j})\, P_{X_1^M|c_{.j}}\right) \\ &= H\left(X_1^M\right)\end{aligned}$$

$$\begin{aligned}\sum_{j=1}^{k_2} \alpha_j H\left(P_{X_1^M|c_{.j}}\right) &= \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j})\, H\left(P_{X_1^M|c_{.j}}\right) \\ &= H\left(X_1^M|X_2^M\right)\end{aligned}$$

On en déduit :

$$\begin{aligned}\lim_{m\to+\infty} \frac{J_{inter}\left(X_2^M\right)}{m} &= \sum_{j=1}^{k_2} P_{X_2^M}(c_{.j})\, H\left(X_1^M|c_{.j}\right) \\ &= H\left(X_1^M\right) - H\left(X_1^M|X_2^M\right) \\ &= MI\left(X_1^M; X_2^M\right)\end{aligned}$$

□

4.9.4 Interprétation asymptotique de typicité

Pour une valeur $v_{.\lambda}$ appartenant au cluster $c_{.\gamma}$ de la variable-partition X_2^M, la typicité est définie comme l'impact moyen sur l'inertie intra-cluster de la retirer de son cluster et de la réaffecter dans un autre cluster de la partition X_2^M.

$$\lim_{m\to+\infty} \frac{T_v(v_{.\lambda})}{m} = -P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || P_{X_1^M|c_{.\gamma}})$$
$$+ \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j}\in X_2^M \\ c_{.j}\neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || \alpha_j P_{X_1^M|c_{.j}} + \alpha_\lambda P_{X_1^M|v_{.\lambda}})$$

où $\alpha_\lambda = \dfrac{P_{X_2}(v_{.\lambda})}{P_{X_2}(v_{.\lambda}) + P_{X_2^M}(c_{.j})}$ et $\alpha_j = \dfrac{P_{X_2}(c_{.j})}{P_{X_2}(v_{.\lambda}) + P_{X_2^M}(c_{.j})}$

Démonstration.

$$\frac{T_v(v_{.\lambda})}{m} = \frac{1}{m(1 - P_{X_2^M}(c_{.\gamma}))} \sum_{\substack{c_{.j}\in X_2^M \\ c_{.j}\neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) (J_{intra}(X_2^M | c_{.\gamma} \setminus v_{.\lambda}, c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M))$$

On a vu que : $\lim\limits_{m\to+\infty} \dfrac{J_{intra}(X_2^M)}{m} = \sum\limits_{j=1}^{k_2} P_{X_2^M}(c_{.j}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.j}\})$

On peut exprimer la typicité en divergences de Jensen-Shannon :

$$\begin{aligned}
\frac{T_v(v_{.\lambda})}{m} &= \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j}\in X_2^M \\ c_{.j}\neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) \left(- \sum_{r=1}^{k_2} P_{X_2^M}(c_{.r}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.r}\}) \right. \\
&\quad \left. + \sum_{r=1}^{k_2} P_{X_2^M}(c_{.r}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.r}, v_{.\lambda} \in c_{.j}, v_{.\lambda} \notin c_{.\gamma}\}) \right) \\
&= \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j}\in X_2^M \\ c_{.j}\neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) \left(- \sum_{r=1}^{k_2} P_{X_2^M}(c_{.r}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.r}\}) \right. \\
&\quad + \sum_{r=1}^{k_2} P_{X_2^M}(c_{.r}) JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.r}\}) \\
&\quad + P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || \alpha_j P_{X_1^M|c_{.j}} + \alpha_\lambda P_{X_1^M|v_{.\lambda}}) \\
&\quad \left. - P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || P_{X_1^M|c_{.\gamma}}) \right)
\end{aligned}$$

$$\frac{T_v(v_{.\lambda})}{m} = \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) \, (\, - P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || P_{X_1^M|c_{.\gamma}})$$

$$+ P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || \alpha_j P_{X_1^M|c_{.j}} + \alpha_\lambda P_{X_1^M|v_{.\lambda}}) \Big)$$

$$= - P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || P_{X_1^M|c_{.\gamma}})$$

$$+ \frac{1}{1 - P_{X_2^M}(c_{.\gamma})} \sum_{\substack{c_{.j} \in X_2^M \\ c_{.j} \neq c_{.\gamma}}} P_{X_2^M}(c_{.j}) P_{X_2}(v_{.\lambda}) KL(P_{X_1^M|v_{.\lambda}} || \alpha_j P_{X_1^M|c_{.j}} + \alpha_\lambda P_{X_1^M|v_{.\lambda}})$$

□

4.9.5 Interprétation asymptotique de l'ajout d'une valeur dans un cluster

Asymptotiquement, la variation d'inertie intra-cluster liée à l'ajout d'une valeur $v_{.\lambda}$ dans un cluster $c_{.j}$ s'interprète comme une divergence de Jensen-Shannon.

$$c^*_{.j} \underset{m\to+\infty}{=} \underset{c_{.j}\in X_2^M}{\operatorname{argmin}} \left[P_{X_2^M}(c_{.j})JS(P_{X_1^M|v_{.\lambda}}, P_{X_1^M|c_{.j}}) - P_{X_2}(v_{.\lambda})JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.j}\})\right]$$

Démonstration. Pour une valeur $v_{.\lambda}$ de la variable X_2, le meilleur cluster d'attribution est celui qui minimise la variation de l'inertie intra-cluster lorsque la valeur $v_{.\lambda}$ lui est ajoutée.

$$c^*_{.j} = \underset{c_{.j}\in X_2^M}{\operatorname{argmin}}(J_{intra}(X_2^M|c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M))$$

L'ajout d'une valeur dans un cluster est équivalent à ajouter une valeur dans son propre cluster et à la changer de cluster.

$$\begin{aligned} J_{intra}(X_2^M|c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M) = &(J_{intra}(X_2^M|c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M|c_{.\lambda} = v_{.\lambda})) \\ &+ (J_{intra}(X_2^M|c_{.\lambda} = v_{.\lambda}) - J_{intra}(X_2^M)) \end{aligned}$$

Un cluster contenant une valeur a une inertie intra-cluster nulle.

$$\begin{aligned} J_{intra}(X_2^M|c_{.\lambda} = v_{.\lambda}) &= (1 - P_{X_2}(v_{.\lambda}))J_{intra}(X_2^M) \\ \Rightarrow J_{intra}(X_2^M|c_{.\lambda} = v_{.\lambda}) - J_{intra}(X_2^M) &= -P_{X_2}(v_{.\lambda})J_{intra}(X_2^M) \end{aligned}$$

En appliquant le théorème de Huygens :

$$J_{intra}(X_2^M|c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M|c_{.\lambda} = v_{.\lambda}) = P_{X_2^M}(c_{.j})JS(P_{X_1^M|v_{.\lambda}}, P_{X_1^M|c_{.j}})$$

On en déduit :

$$J_{intra}(X_2^M|c_{.j} \cup v_{.\lambda}) - J_{intra}(X_2^M) = P_{X_2^M}(c_{.j})JS(P_{X_1^M|v_{.\lambda}}, P_{X_1^M|c_{.j}}) - P_{X_2}(v_{.\lambda})J_{intra}(X_2^M)$$

Or seul l'inertie intra-cluster locale au cluster $c_{.j}$ a un impact sur le choix du cluster. D'où :

$$c^*_{.j} \underset{m\to+\infty}{=} \underset{c_{.j}\in X_2^M}{\operatorname{argmin}} \left[P_{X_2^M}(c_{.j})JS(P_{X_1^M|v_{.\lambda}}, P_{X_1^M|c_{.j}}) - P_{X_2}(v_{.\lambda})JS(\{P_{X_1^M|v_{.i}}, \forall v_{.i} \in c_{.j}\})\right]$$

5

Applications

On a vu dans les précédents chapitres que le co-clustering permet de traiter de nombreux problèmes : graphes statiques ou temporels, courbes ou données relationnelles plus complexes. L'approche MODL permet d'obtenir des résultats fins et pertinents. Un post-traitement ainsi que des indicateurs ont été mis en place afin de pouvoir explorer les résultats du co-clustering et d'en extraire l'information la plus utile et la plus pertinente.

Dans ce chapitre, des études variées et détaillées sont présentées pour l'analyse de compte-rendus d'appels, en exploitant au mieux les informations extraites par une seule méthode : le co-clustering. Une méthodologie d'analyse exploratoire des données est également proposée.

5.1 Préliminaires

Dans un premier temps, les données utilisées ainsi que les études réalisées sont présentées. Ces données ont été collectées en Côte d'Ivoire entre Décembre 2011 et Avril 2012 sur le réseau de l'opérateur Orange. Elles ont fait l'objet de traitements préalables d'anonymisation. Afin de comprendre la démarche de l'étude, la méthodologie d'analyse est également expliquée dans cette partie.

5.1.1 Descriptif des données et études menées

Étude des communications mobiles Ivoiriennes les données collectées sont des appels mobiles passés en Côte d'Ivoire entre clients Orange dans la période allant du 1^{er} décembre 2011 au 28 Avril 2012. Ces données sont issues du challenge D4D (Data For Development, (Blondel *et al.*, 2012)). Le nombre d'enregistrements est d'environ 471 millions d'appels, décrits par les variables suivantes :

- antenne émettrice (variable nominale avec 1 214 modalités),
- antenne réceptrice (variable nominale avec 1 216 modalités),
- heure de l'appel (précision à l'heure),
- la date (du 1^{er} décembre 2011 au 28 Avril 2012).

À partir de ces données, nous proposons plusieurs études :

1. *étude du réseau d'appels entre antennes,* les variables utilisées sont les antennes émettrices et réceptrices. On a donc affaire à un multigraphe orienté où les nœuds sont les antennes et les arcs les appels ;

2. *étude du trafic sortant en fonction de la date,* les variables utilisées sont les antennes émettrices et le nombre de jours écoulés depuis le début de la collecte des données (date de l'appel). Il s'agit du cas d'un biclustering avec une variable nominale et une variable continue ;

3. *étude du trafic sortant en fonction du jour de la semaine et de l'heure de la journée,* les variables utilisées sont les antennes émettrices, le jour de la semaine (déduit de la date et traité comme une variable continue) et l'heure de la journée. Il s'agit du cas d'un triclustering avec une variable nominale et deux variables continues.

Cette étude est détaillée en section 5.2.

Étude des communications mobiles de la Côte d'Ivoire vers l'international une seconde base n'étant pas issue du challenge D4D est également

étudiée. Il s'agit des communications (appels, SMS...) entre le réseau mobile Orange de la Côte d'Ivoire et l'international. Le trafic des mobiles vers l'international est constitué d'environ 22 millions de communications alors que le trafic de l'international vers les mobiles Ivoiriens représente environ 13 millions de communications. Les données disponibles sont les suivantes :

- antenne émettrice (variable nominale avec 1 214 modalités) ou pays d'origine de l'appel,
- antenne réceptrice (variable nominale avec 1 216 modalités) ou pays de destination de l'appel,
- type de service (variable nominale. Ex : appel, SMS...),
- date (du 1^{er} Février au 31 Mars 2012),
- heure de l'appel (précision à la minute).

Les études suivantes sont proposées :

1. *étude des communications émises depuis le réseau mobile Ivoirien vers l'international*, les variables utilisées sont les antennes émettrices et les pays de destination de l'appel. Il s'agit donc d'un graphe biparti que l'on traite par un biclustering de deux variables nominales. Nous réalisons également l'étude des communications de l'international vers les mobiles Ivoiriens ;

2. *étude des communications émises depuis l'international vers le réseau mobile Ivoirien en fonction de l'heure de la journée*, les variables utilisées sont les antennes réceptrices, les pays émetteurs des appels et l'heure de la journée. Il s'agit du cas d'un triclustering avec deux variables nominales et une variable continue, assimilable à un graphe biparti temporel dont l'un des ensembles de nœuds correspond aux antennes réceptrices et l'autre aux pays émetteurs des appels, ces derniers étant modélisés par des arcs estampillés de l'heure de la communication ;

3. *étude des communications émises depuis l'international vers le réseau mobile Ivoirien, en fonction de l'heure de la journée, du jour de la semaine et du type de service*, on a donc cinq variables étudiées : trois sont nominales et deux continues. Il s'agit d'un pentaclustering.

Cette étude est détaillée en section 5.3.

Études de mobilité Enfin une dernière base de données est étudiée, également issue du challenge D4D. Il s'agit de traces de mobilité des clients du réseau. Pendant une période de deux semaines, les antennes utilisées par 50 000 clients Orange sont enregistrées à chacune de leur utilisation du réseau. Le

nombre d'utilisations collectées est de 55 millions. On dispose des variables suivantes :

- identifiant anonymisé de l'utilisateur (variable nominale avec 50 000 modalités),
- antenne de connexion (variable nominale avec 1 214 modalités),
- heure de l'appel (précision à la minute),
- la date (10 Décembre 2012 au 24 Décembre 2012).

Grâce à ces données, nous pouvons assimiler l'utilisation d'une antenne à un point de présence dans un réseau et l'identifiant de l'utilisateur à une trajectoire. De là, on décide de mener trois études :

1. *étude des déplacements des usagers*, les variables utilisées sont les identifiants et les antennes. On réalise un biclustering entre deux variables nominales ;

2. *détection des utilisateurs les plus mobiles*, on construit, à partir des données, des courbes d'utilisation des antennes de chaque usagers afin de détecter les individus les plus mobiles. Pour cela, nous réalisons un clustering de courbes, c'est-à-dire un triclustering avec deux variables continues et une variable nominale ;

3. *étude des déplacements des usagers en fonction du jour de la semaine et de l'heure de la journée*, les variables utilisées sont les identifiants, les antennres, le jour de la semaine et l'heure de la journée. On réalise un tetraclustering entre trois variables nominales et une variable continue.

Cette étude est détaillée en section 5.4.

5.1.2 Méthodologie d'analyse

Cette section a pour objectif d'introduire une méthodologie d'analyse générique utilisant les différents éléments d'analyse exploratoire introduits dans le chapitre 4. Nous considérons ici avoir modélisé le problème et appliqué l'approche MODL. Nous disposons donc d'une grille de co-clustering dont nous souhaitons extraire une information exploitable et utile afin d'analyser les clusters les plus intéressants. Les étapes de l'analyse exploratoire sont détaillées sur la figure 5.1.

Afin de mieux comprendre ce processus nous détaillons les différentes étapes et mesures nécessaires à leur mise en place :

1. **le co-clustering le plus fin :** il s'agit du co-clustering obtenu après application de l'approche MODL. On peut analyser les première statistiques descriptives des partitions :
   - *le nombre de clusters* : on l'analyse pour chacune des partitions et en fonction de la finesse du partitionnement des variables descriptives des données, on décide d'exploiter les résultats en l'état ou non ;
   - *le taux de dispersion des clusters* : cette mesure est définie comme l'inertie inter-clusters normalisée d'une partition. Elle est comprise entre 0 et 1. Plus elle est importante, plus les lois de probabilités des partitions conditionnelles aux clusters sont dissimilaires. Asymptotiquement, le taux de dispersion d'une partition est l'information mutuelle entre la partition étudiée et le produit cartésien des autres partitions. On utilise cette mesure pour évaluer la corrélation de la partition d'une variable avec la partition conjointe des autres variables. Voir section 4.4.1 ;
   - *le taux de dispersion des valeurs* : il s'agit de la dispersion moyenne des valeurs dans leur cluster. On le définit comme l'inertie intra-cluster normalisée. Cette mesure est également comprise entre 0 et 1. Plus elle est faible, plus les clusters sont compacts, c'est-à-dire qu'ils groupent des valeurs très similaires les unes par rapport aux autres. On se sert de cette mesure comme un outil d'analyse de la qualité des clusters. Voir section 4.4.2 ;

2. **le co-clustering simplifié :** la simplification du co-clustering est une étape facultative qui est opérée si la partition des valeurs des variables est trop fine pour que des analyses puissent être menées. La simplification se fait hiérarchiquement par fusions successives des clusters, en maîtrisant la dégradation du modèle. Deux démarches de simplifications peuvent être envisagées :
   - *une simplification globale des résultats :* les clusters sont fusionnés successivement sur toutes les partitions. À chaque étape du post-traitement hiérarchique, les deux clusters fusionnés sont ceux qui minimisent la dégradation du modèle de co-clustering, parmi l'ensemble des clusters de toutes les partitions. La dégradation est quantifiée par une mesure d'informativité correspondant au critère MODL normalisé par Min-

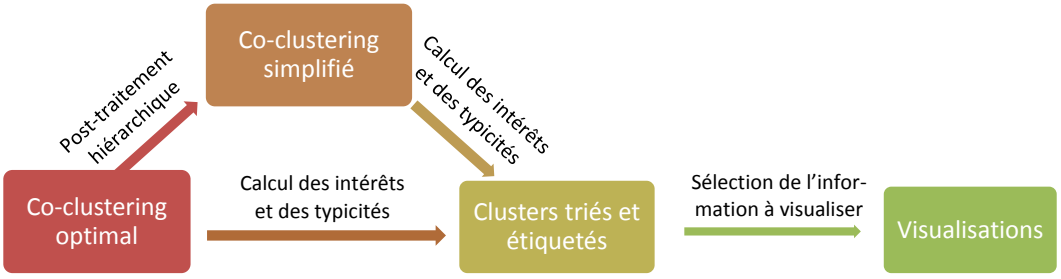


FIGURE 5.1 – Processus de l'analyse exploratoire des grilles de co-clustering

Max, avec pour minimum la valeur du critère du modèle nul (un seul co-cluster) et pour maximum la valeur du critère du modèle optimal. Voir section 4.3.2 ;

- *une simplification indépendante des partitions :* les clusters sont fusionnés successivement sur une seule partition (l'autre étant maintenue à son niveau optimal) de manière à dégrader au minimum le modèle. La dégradation est quantifiée par une mesure d'informativité correspondant à l'inertie inter-clusters, normalisée par la valeur de l'inertie inter-clusters de la partition à son niveau optimal. Voir section 4.3.2 ;

3. **clusters triés et étiquetés :** une fois que le niveau de précision de la grille est satisfaisant pour envisager une exploration des résultats, les clusters sont triés par intérêt décroissant et étiquetés par la valeur la plus typique du cluster ;

   - *l'intérêt des clusters :* il s'agit mesurer l'impact des clusters sur l'inertie inter-clusters de leur partition. Asymptotiquement, l'intérêt d'un cluster est un compromis entre sa probabilité d'apparition dans les données et la divergence de Jensen-Shannon entre la loi jointe des variables-partitions et la loi d'une variable-partition conditionnellement à un cluster de l'autre variable-partition. Un cluster intéressant est donc un cluster très différent des autres mais groupant des valeurs suffisamment observées dans les données pour être significatif. Voir section 4.5.1 ;
   - *la typicité des valeurs :* une valeur est dite typique si elle représente bien son cluster. La typicité est mesurée comme le coût moyen de réaffectation d'une valeur dans les clusters auxquels elle n'appartient pas. Plus il est coûteux de déplacer la valeur de son cluster, plus elle en est une bonne représentante. Voir section 4.5.2 ;

4. **la visualisation :** plusieurs vues sont proposées dans cette étude. Ces visualisations apparaissent sous forme de matrices en deux dimensions, présentant les interactions entre deux partitions ; ou alors sous forme de cartes puisque l'étude concerne, entre autres, une segmentation géographique ;
   - *l'information mutuelle* : elle mesure la corrélation entre les partitions de deux variables. On s'intéresse ici à la contribution de chaque bicluster à l'information mutuelle. Un bicluster avec une contribution positive (resp. négative) à l'information mutuelle met en en évidence un excès (resp. un déficit) d'interactions entre les deux clusters formant le bicluster, par rapport à la quantité d'interactions attendue en cas d'indépendance des deux partitions. Lorsque la contribution est nulle, cela signifie qu'il y a la quantité d'interactions attendue ou alors une quantité très faible ou nulle. Voir section 4.7.1 ;

– *le contraste* : le contraste est une mesure définie lorsqu'il y a plus de deux partitions à étudier. Basée sur l'information mutuelle, le contraste permet de calculer les excès et déficits d'interactions entre les biclusters visualisés et un co-cluster formé par le produit cartésien des clusters sélectionnés sur chacune des partitions des autres variables. Voir section 4.7.2 ;

5. **interprétation des résultats :** des interprétations socio-économique des résultats sont proposées. Ces interprétations ont été menées conjointement avec le Professeur Akindès de l'Université de Bouaké dans le cadre d'un projet industriel, interne à Orange.

5.2 Étude des communications mobiles Ivoiriennes.

La première base de données étudiée est un compte-rendu d'appels passés en Côte d'Ivoire entre clients Orange dans la période allant du 1^{er} décembre 2011 au 28 Avril 2012. Ces données sont agrégées par antennes. Trois études sont proposées dans cette section : une analyse des appels entre antennes, une analyse du trafic sortant en fonction de la date et enfin une analyse du trafic sortant en fonction du jour de la semaine et de l'heure de la journée.

5.2.1 Étude des communications entre antennes

On s'intéresse dans un premier temps au graphe de communications inter-antennes. Le graphe est composé de 1 216 nœuds (1 214 antennes émettrices et 1 216 réceptrices) et de 471 millions d'arcs. Le graphe est traité comme un multigraphe biparti, on cherche donc deux partitions : une partition des antennes émettrices et une partition des antennes réceptrices.

	Partitions des antennes émettrices	**Partitions des antennes réceptrices**
Nombre de clusters	1150	1150
Inertie inter-clusters	$1{,}14.10^9$	$1{,}14.10^9$
Taux de dispersion des clusters	39,40%	39,33%
Inertie intra-cluster	$1{,}24.10^6$	$1{,}21.10^6$
Taux de dispersion des valeurs	14,09%	14,04%

TABLE 5.1 – Caractéristiques des partitions

Étude des clusters Le tableau 5.1 donne les caractéristiques relatives aux partitions des antennes émettrices et réceptrices. On remarque que le nombre de clusters est le même pour les deux partitions. D'ailleurs, les deux partitions sont identiques à l'exception de deux clusters contenant chacun une des deux antennes réceptrices et non-émettrices et dont le trafic est très peu significatif (un et onze appels reçus). La valeur du taux de dispersion des clusters (inertie inter-clusters normalisée) approche les 40% pour les deux partitions, ce qui nous indique que les clusters sont bien séparés : chaque cluster émet (resp. reçoit) des appels vers (resp. depuis) les mêmes clusters. En analysant les résultats de plus près, une structure très homophile apparaît, c'est-à-dire que les clusters ont tendance à s'appeler eux-même.

On observe un taux de dispersion dans les clusters contenant plusieurs antennes d'environ 14%, ce qui peut sembler élevé. Cette valeur due au fait que les rares clusters groupant plusieurs antennes sont relativement dispersés. On peut les qualifier de clusters « poubelles » : ces clusters sont parmi les moins intéressants selon la mesure d'intérêt introduite dans la section 4.5.1. En analysant la valeur de l'inertie intra-cluster, on se rend compte que celle-ci est environ mille fois inférieure à l'inertie inter-clusters, ce qui montre que les clusters sont très différents les uns des autres et groupent des antennes très similaires.

Un modèle de co-clustering de cette finesse est finalement peu utile car il ne compresse presque pas les données d'origine. On applique donc la simplification hiérarchique proposée dans la section 4.3.2. De manière à pouvoir interpréter la mesure à l'échelle nationale, on choisit de conserver 60% de l'information apportée par le modèle, ce qui correspond à conserver 20 clusters d'antennes émettrices et réceptrices (voir figure 5.2). Notons que la simplification se fait séquentiellement sur les antennes source et cible. La symétrie est conservée à tous les niveau de la hiérarchie. Même à ce niveau de précision, la symétrie entre les deux partitions est conservée. On choisit donc de se focaliser sur la partition des antennes émettrices.

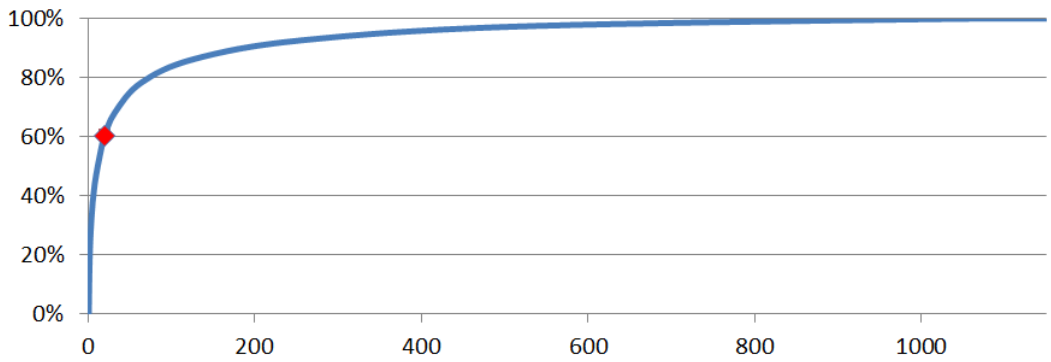


FIGURE 5.2 – Courbe de l'information du modèle de co-clustering en fonction du nombre de clusters conservés lors du post-traitement hiérarchique ascendant

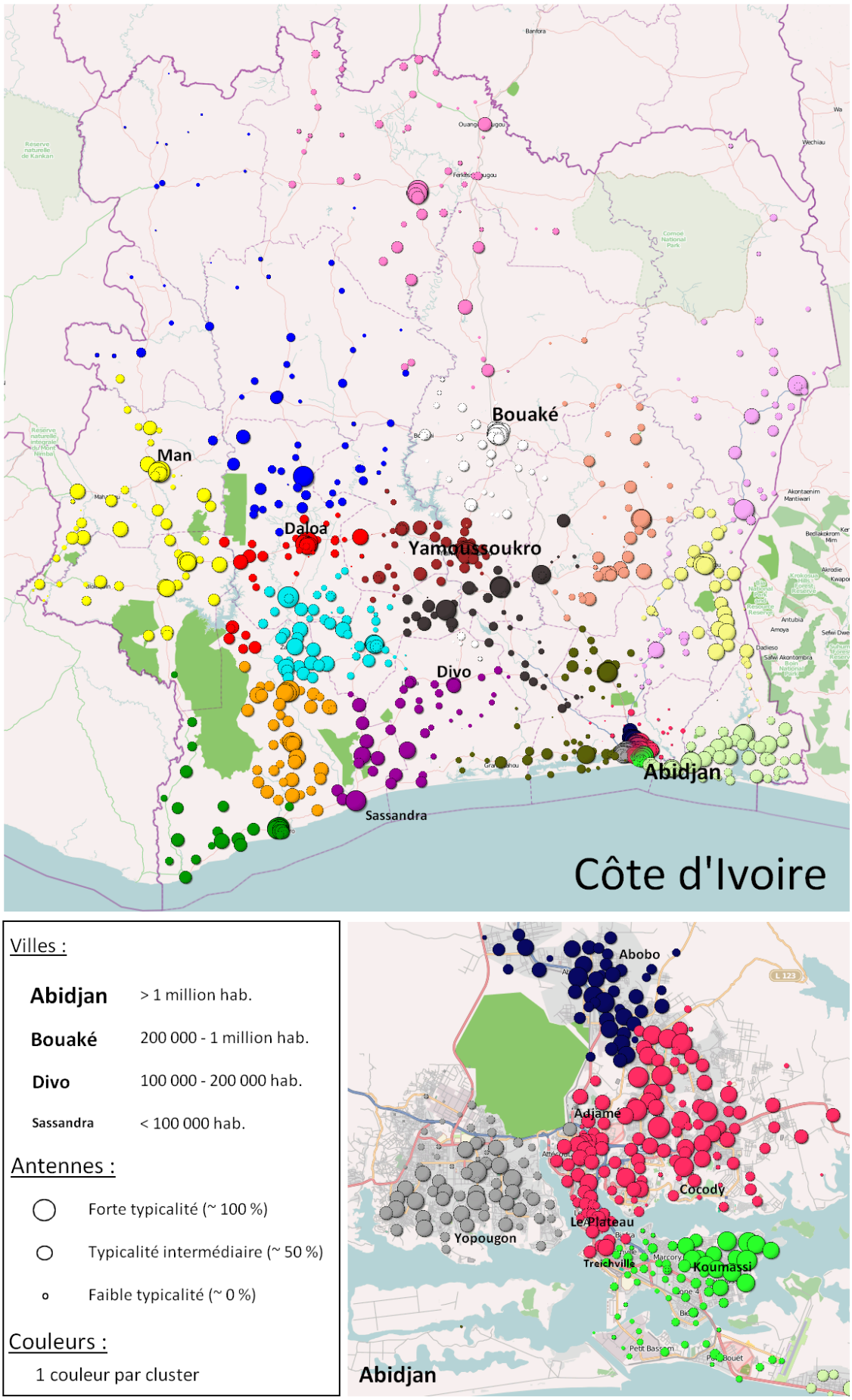


FIGURE 5.3 – Projection des vingt clusters sur une carte de la Côte d'Ivoire

Id Cluster	Intérêt	Antennes par cluster	Trafic émis	Couleur sur la carte
Abj.-Centre	1	14,50%	17,92%	fuschia (Sud-Est)
Man	0,4557	4,86%	4,90%	jaune (Ouest)
Yamoussoukro	0,4195	5,82%	7,05%	marron (Centre)
Soubré	0,4143	4,56%	6,90%	orange (Sud-Ouest)
Bonoua	0,3964	6,36%	6,84%	vert pâle (Sud-Est)
Issia	0,3963	7,52%	7,52%	cyan (Centre-Ouest)
San Pédro	0,3872	3,94%	7,19%	vert foncé (Sud-Ouest)
Daloa	0,3755	6,01%	8,29%	rouge (Centre)
Bouaké	0,3638	7,99%	10,83%	blanc (Centre)
Abengourou	0,3636	6,95%	9,17%	jaune pâle (Est)
Abj.-Yopougon	0,3540	6,84%	7,17%	gris (Sud-Est)
Vavoua	0,3385	6,01%	3,69%	bleu (Nord-Ouest)
Abj.-Abobo	0,3192	4,20%	6,34%	bleu marine (Sud-Est)
Korhogo	0,2839	4,70%	3,14%	rose (Nord)
Agnibilékrou	0,2805	6,01%	3,69%	mauve pâle (Est)
Oumé	0,2694	3,95%	4,05%	noir (Centre)
Sassandra	0,2480	3,54%	3,33%	violet (Sud)
Abj.-Koumassi	0,2378	6,67%	2,93%	vert fluo (Sud-Est)
Daoukro	0,2260	3,62%	3,02%	saumon (Centre-Est)
Agboville	0,1963	4,20%	2,64%	kaki (Sud)

TABLE 5.2 – Vingt clusters d'antennes de Côte d'Ivoire et leur caractéristiques (Abj. = Abidjan).

Afin de faciliter l'interprétation du clustering, on calcule l'intérêt de chacun des clusters, ainsi que la typicité des antennes. On choisit d'ailleurs d'attribuer un label au cluster correspondant à la localité où se trouve l'antenne la plus typique. Les clusters et leurs caractéristiques sont détaillés dans le tableau 5.2. Par ailleurs, les clusters sont projetés sur une carte. Les antennes sont représentées par des points sur la carte de la figure 5.3, un point est coloré en fonction de son cluster d'appartenance. La taille du point est proportionnelle au produit de la typicité de l'antenne et de l'intérêt de son cluster.

La première chose qui apparaît lorsque qu'on regarde la carte est la corrélation entre le regroupement des antennes et leur situation géographique. Cette observation couplée à l'homophilie (majorité d'observations sur la diagonale) de la structure de biclustering montre que les usagers appellent principalement dans la région où ils se trouvent. On parle donc de « régions téléphoniques » pour désigner les clusters. Les régions sont de tailles similaires et correspondent grosso modo aux subdivisions administratives du pays, à l'exception de la ville d'Abidjan qui est divisée en quatre clusters. Ce découpage d'Abidjan s'explique par la très forte densité d'antennes dans la ville (32,21% des antennes et 34,36%

du trafic).

Pour chacune de ces régions, l'antenne avec la typicité la plus forte est utilisée en tant qu'étiquette du cluster. Les clusters identifiés grâce au nom de la commune où se trouve leur antenne la plus typique sont présentés dans la table 5.2. On se rend compte que les antennes typiques se situent dans les principales villes du pays (Bouaké, Daloa, Yamoussoukro...). On retrouve ici un phénomène déjà observé (Blondel *et al.*, 2010; Guigourès et Boullé, 2011) de correspondance entre la zone d'influence des métropoles avec les clusters d'antennes. Certaines exceptions sont à noter parmi lesquelles la région téléphonique de Sassandra (40 000 habitants) dans laquelle se trouve la ville de Divo (6^e ville du pays avec 185 000 habitants). Les antennes de la ville de Divo ont une typicité 40% plus faible que l'antenne la plus typique située à Sassandra. Ce qui signifie qu'affecter les antennes de Divo à un autre cluster ne serait pas si coûteux sur le critère, et donc que le trafic provenant de cette ville est significatif vers les autres clusters. Ce résultat n'est pas étonnant quand on sait que Divo est une ville connaissant une très forte croissance démographique (la population a doublé en 20 ans) liée à des migrations internes au pays (Gnabéli, 2008). D'autre part la région autour de Divo a connu d'importants changements ces dernières années. L'agriculture vivrière de la région a été remplacée par une culture intensive de l'hévéa, drainant d'importantes populations de saisonniers venus des autres régions du pays.

En s'intéressant à Abidjan, on s'aperçoit que la ville est divisée en quatre clusters très corrélés avec la géographie sociale de la ville :

- le cluster couvrant le centre-ville est celui qui maximise la mesure d'intérêt mais également celui d'où provient près de 18% du trafic du pays. Les quartiers centraux d'Abidjan se divisent en trois principaux quartiers : le Plateau (quartier des affaires), Adjamé (quartier commerçant et hub ferroviaire et routier du pays) et enfin Cocody (quartier des ambassades, de l'université et zones résidentielles aisées) ;
- les clusters d'Abobo et de Yopougon ont tous les deux le même profil bien qu'ils constituent deux différents clusters. Il s'agit des quartiers résidentiels défavorisés de la ville ;
- les quartiers sud de Koumassi, Marcory et Port-Bouët sont couverts par le même cluster. Ces quartiers sont partagés entre zones portuaires et quartiers résidentiels. Notons que le quartier de Treichville est partagé en deux : l'extrême Nord, connu pour sa vie nocturne, est rattachée au cluster du centre-ville alors que la partie Sud, zone portuaire, est rattachée au cluster voisin de Koumassi.

La mesure d'intérêt est un compromis entre le poids du cluster (nombre d'appels) et la divergence de la distribution d'appels provenant de ses antennes par rapport à la distribution moyenne de toutes les antennes. Ainsi, lorsque les clusters sont très déséquilibrés, les petits clusters ont besoin de diverger de manière très significative de la moyenne des données pour présenter un fort intérêt. C'est pour cette raison que le pourcentage de trafic sortant est également

présenté dans le tableau. Le cluster de Man a une forte valeur d'intérêt par rapport au nombre d'appels qui y sont émis. Pour le cluster de Bouaké, c'est l'inverse. On a vu précédemment que la forte inertie inter-clusters était la conséquence d'une tendance pour les clusters à grouper des antennes ayant un grand nombre d'appels entre elles. Le cluster de Man est donc très centré sur lui même, ce qui fait de Man une ville avec une forte influence régionale, alors que le cluster de Bouaké a une tendance plus importante à communiquer avec le reste du pays, ce qui en fait une métropole au rayonnement régional et national.

Étude du trafic entre clusters Dans le chapitre 4, la contribution à l'information mutuelle a été introduite afin de visualiser les excès et les déficits d'observations dans les co-clusters par rapport aux effectifs attendus. Nous proposons ici d'étudier cette contribution sur les antennes, ce qui revient à analyser les excès et les déficits d'appels entre clusters d'antennes.

Peu importe le niveau de précision du modèle, on observe des excès d'appels des clusters vers eux-mêmes et des excès ou des déficit de trafic bien plus faibles entre les clusters. L'intérêt d'étudier le trafic interne aux clusters étant assez limité à cette échelle d'analyse, on s'intéresse au trafic entre les clusters. La carte 5.4 présente les excès d'appels entre clusters pour un modèle simplifié à 355 clusters sur les deux partitions, conservant 95% de l'information du modèle le plus fin. Les traits rouges sur la carte représentent les excès de trafic, leur opacité est fonction de la contribution à l'information mutuelle de la paire de clusters, et l'épaisseur du trait au volume d'appels.

Les métropoles apparaissent clairement sur la carte à l'image des villes de Bouaké, Man ou San Pédro. Ces villes sont parmi les plus grandes du pays, sont des capitales administratives et leur rayonnement à l'échelle régionale est très net sur la carte. Le cas de Bouaké est intéressant. Bien qu'elle ne soit pas la capitale administrative du pays, elle a un plus grand rayonnement national que Yamoussoukro, la capitale. Yamoussoukro est une ville nouvelle et environ deux fois moins peuplée que Bouaké. De plus, l'activité économique ne s'y concentre pas, ce qui peut expliquer le phénomène observé.

Les excès de trafic se font la plupart du temps entre les campagnes et la ville la plus proche et rarement entre villes. Le Centre-Ouest du pays, autour de Soubré, échappe à ce phénomène et les excès de trafic s'observent entre tous les clusters locaux. On ne voit pas réellement de métropole influente émerger. Ces zones ne sont pourtant pas plus densément peuplées que le reste du pays. Cependant elles correspondent à des régions où les déplacements de populations sont importants.

Enfin, l'excès de trafic téléphonique entre Abidjan et le reste du pays n'est pas significatif, si ce n'est le long de la côte. En se penchant sur la ville elle-même, on observe d'importants excès à l'intérieur des quartiers mais très peu en dehors. Ce phénomène est marqué pour Yopougon (à l'ouest) et Koumassi (au Sud). On observe par contre un excès d'appels entre les quartiers du Plateau

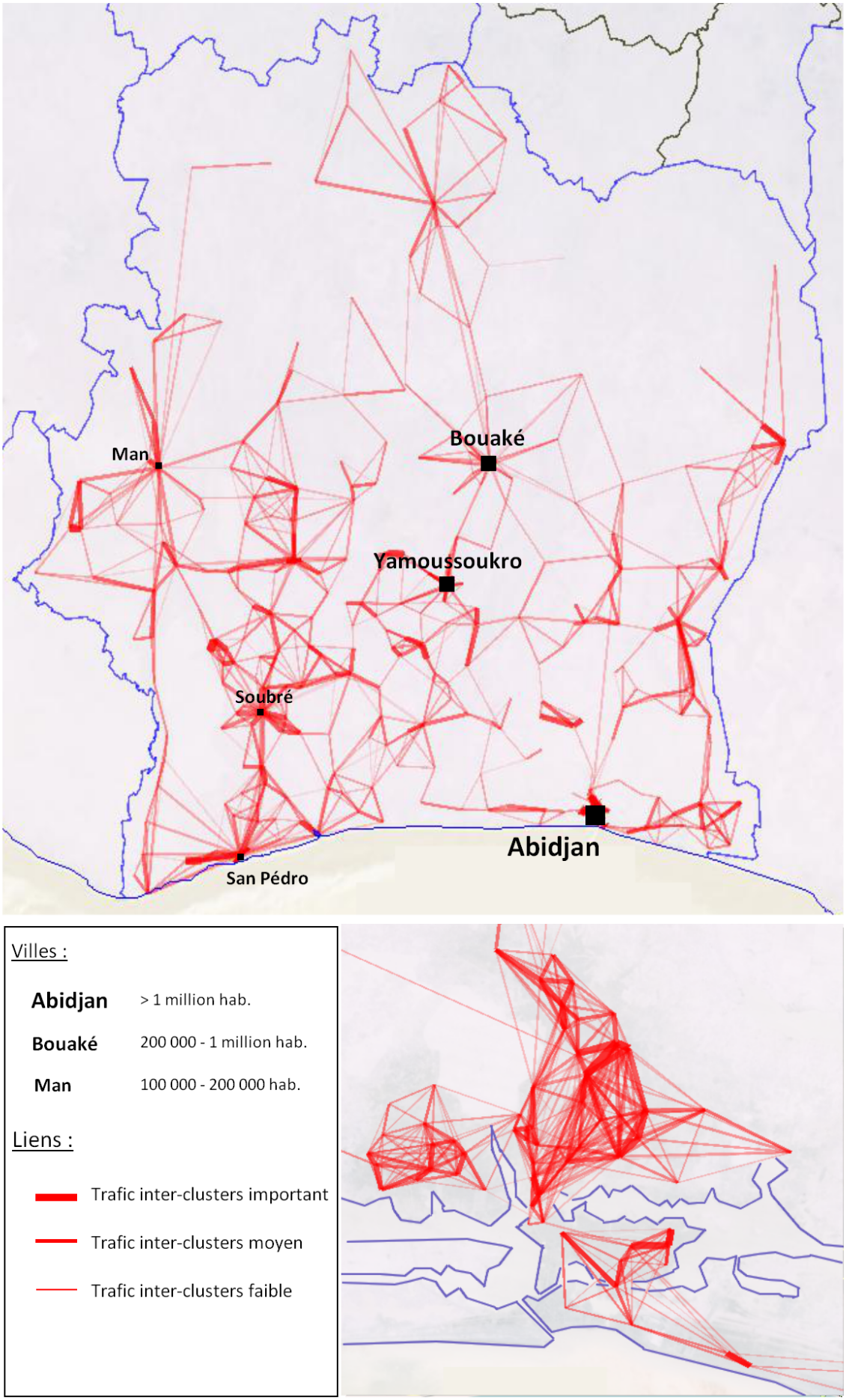


FIGURE 5.4 – Excès d'appels projetés sur une carte de la Côte d'Ivoire et d'Abidjan.

et de Cocody, ce qui peut s'expliquer par la nature socio-économique de ces zones : le Plateau est le quartier des affaires et Cocody le quartier où vivent les classes aisées. Le quartier Nord d'Abobo est populaire mais communique malgré tout avec les quartiers centraux. Ce quartier ne communique cependant pas particulièrement avec Yopougon qui possède pourtant le même profil socio-économique, ce qui souligne une fois de plus la corrélation entre la géographie de la ville et la distribution des appels.

5.2.2 Étude des communications émises en fonction de la date

Dans un second temps, on s'intéresse à la répartition des appels sur la période de temps étudiée, c'est-à-dire du 1^{er} Décembre 2011 au 28 Avril 2012. Pour cela, on réalise un biclustering des antennes émettrices (variable nominale) et du nombre de jours écoulés depuis le début de la collecte des données (variable continue), de manière à obtenir une discrétisation temporelle et une segmentation des antennes en clusters, dans lesquels le trafic sortant se répartit de manière identique sur les périodes de l'année.

Triclustering vs. biclustering. Pourquoi ne pas utiliser le modèle de graphe temporel introduit dans la section 3.2 ? On pourrait trouver plus logique d'étendre l'étude précédente du graphe de communications inter-antennes au cas d'un graphe temporel, en y ajoutant la variable date. L'étude du graphe inter-antennes a montré une corrélation très forte entre les partitions des antennes émettrices et réceptrices. Le taux de dispersion des clusters obtenu, de 40%, et l'observation des interactions entre clusters fait apparaître une structure diagonale et des corrélations très fortes entre les partitions. Avec trois variables, la partition conjointe des antennes est si informative pour le modèle, qu'aucune structure temporelle n'émerge malgré la volumétrie des données. En effet, le partitionnement des antennes fait croître la vraisemblance du modèle de manière plus importante que la discrétisation du temps. La méthode étant régularisée, jamais la structure temporelle ne parvient à être retrouvée. On fait donc l'hypothèse que les variables antennes émettrices et réceptrices sont totalement corrélées et n'utiliser qu'une des deux pour notre étude : les antennes émettrices.

Étude des clusters d'antennes. En appliquant l'approche MODL, on obtient 1 051 clusters d'antennes (pour 1 214 antennes émettrices) et 140 intervalles de temps, soit le nombre total de jours étudiés pour ces données. La valeur du taux de dispersion des clusters est de 5,38%, donc bien inférieure à l'inertie obtenue dans l'étude du trafic inter-antennes, ce qui nous conforte dans le choix de ne pas prendre en compte les antennes réceptrices.

La grille est trop fine pour être analysée directement avec ce niveau de précision. Cette finesse de la partition des antennes s'explique par le volume

Id	Intérêt	Antennes par cluster	Trafic émis	Commentaire
1	1	26,61%	29,30%	Données complètes
2	0,6320	35,09%	44,93%	Données complètes
3	0,0824	9,31%	8,06%	Données complètes
4	0,0616	3,87%	4,62%	Manquant du 16/02 au 23/03/2012[1]
5	0,0591	4,94%	4,24%	Données complètes
6	0,0493	4,28%	3,94%	Manquant du 27/02 au 23/03/2012[1]
7	0,0279	3,29%	2,17%	Manquant du 28/01 au 23/03/2012[1]
8	0,0023	2,14%	0,62%	Manquant du 14/12/2011 au 23/03/2012[1]
9	0,0015	1,32%	0,63%	Manquant du 13/01 au 23/03/2012[1]
10	0,0002	9,14%	1,47%	Début de collecte le 23/03/2012[2]

TABLE 5.3 – Dix clusters d'antennes de Côte d'Ivoire et leurs caractéristiques

des données et par la précision de l'enregistrement des comptes-rendus : il y a en moyenne près de 3 000 appels par jour et par antenne, ce qui permet de les différencier dans cette étude temporelle. On choisit donc de réduire le nombre de clusters sur les deux partitions, tout en conservant 80% de l'information apportée par le modèle. À ce niveau de précision, on a 10 clusters d'antennes émettrices et 20 intervalles de temps. On calcule alors la mesure d'intérêt pour les clusters d'antennes. Le tableau 5.3 donne ces valeurs d'intérêt ainsi que le pourcentage de trafic émis et d'antennes du cluster.

On s'aperçoit que le cluster 10, malgré un trafic significatif et un nombre d'antennes important, a un intérêt très faible (environ 0,0002). Ce cluster correspond à des antennes réparties sur l'ensemble du territoire (antennes jaunes sur la carte de la figure 5.5) dont l'activité démarre à la fin Mars 2012 dans nos données. D'autre part, certaines antennes se retrouvent regroupées car elles ont connu des périodes d'inactivité importantes, supérieures à deux semaines. De manière générale, la valeur d'intérêt décroît avec la durée de la panne, sauf pour le cluster 9. Il est intéressant d'observer que ces clusters sont géographiquement corrélés (voir carte de la figure 5.5). Une hypothèse est que ces antennes sont tombées en panne à la même date, dans des secteurs délimités et très localisés.

Le problème de ces données est qu'elles provoquent des discrétisations artificielles liées aux données manquantes. La partition des antennes étant obtenue conjointement à la discrétisation, ces phénomènes impactent également les clusters d'antennes : le tableau 5.3 montre clairement une corrélation entre

1. Ces données manquantes peuvent être liées à des pannes ou des problèmes de collectes de données.

2. L'apparition de ces nouvelles antennes peut être expliquée par une mise en service de nouveaux sites ou par un problème de collectes de données.

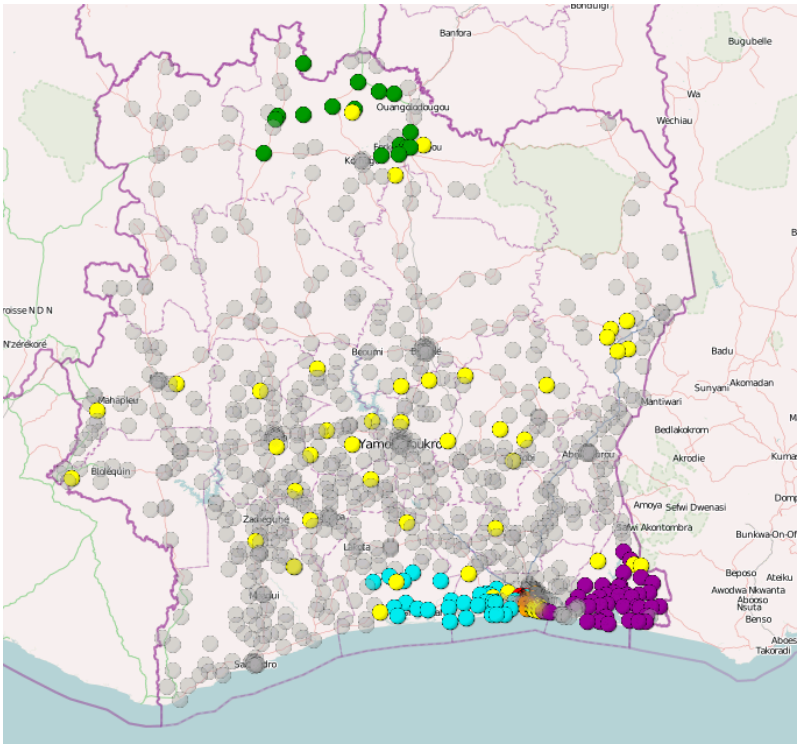


FIGURE 5.5 – Clusters projetés sur une carte de la Côte d'Ivoire. Les clusters colorés sont les clusters dont les antennes ont des périodes d'inactivité. Il y a une couleur par cluster. Les clusters gris sont les clusters d'antennes dont le trafic est complet sur la durée de l'étude.

les durées d'inactivité et les regroupements d'antennes, les antennes groupées ont toutes des enregistrements manquants aux mêmes périodes. On choisit de ne pas étudier les données au-delà du 27 Février. De plus, toutes les antennes ayant connu des pannes avant le 27 Février sont retirés de la base. En procédant ainsi, on perd 38,02% des appels, 19,77% des antennes et on réduit de 32,86% la durée de l'étude.

L'approche MODL est réappliquée sur la nouvelle base de données filtrée. Nous obtenons des partitions très fines à l'image de celles que nous avions obtenues avant nettoyage des données. En regardant la répartition des clusters sur le territoire, la forte corrélation géographique n'est pas observée comme dans l'étude du trafic entre antennes, cependant les clusters se limitent à des secteurs bien définis du pays : les clusters sont dispersés soit dans l'Est soit dans l'Ouest du pays. En appliquant le post-traitement hiérarchique de la section 4.3.2 et en réduisant le nombre de clusters d'antennes à deux (et quatre intervalles de temps, soit 65% d'information conservée), le pays est coupé en deux par un axe Nord-Sud passant à l'Ouest de Yamoussoukro.

Études des excès et déficits de trafic. Afin de mieux comprendre ce phénomène, nous allons étudier les excès et les déficits d'appels en fonction de la

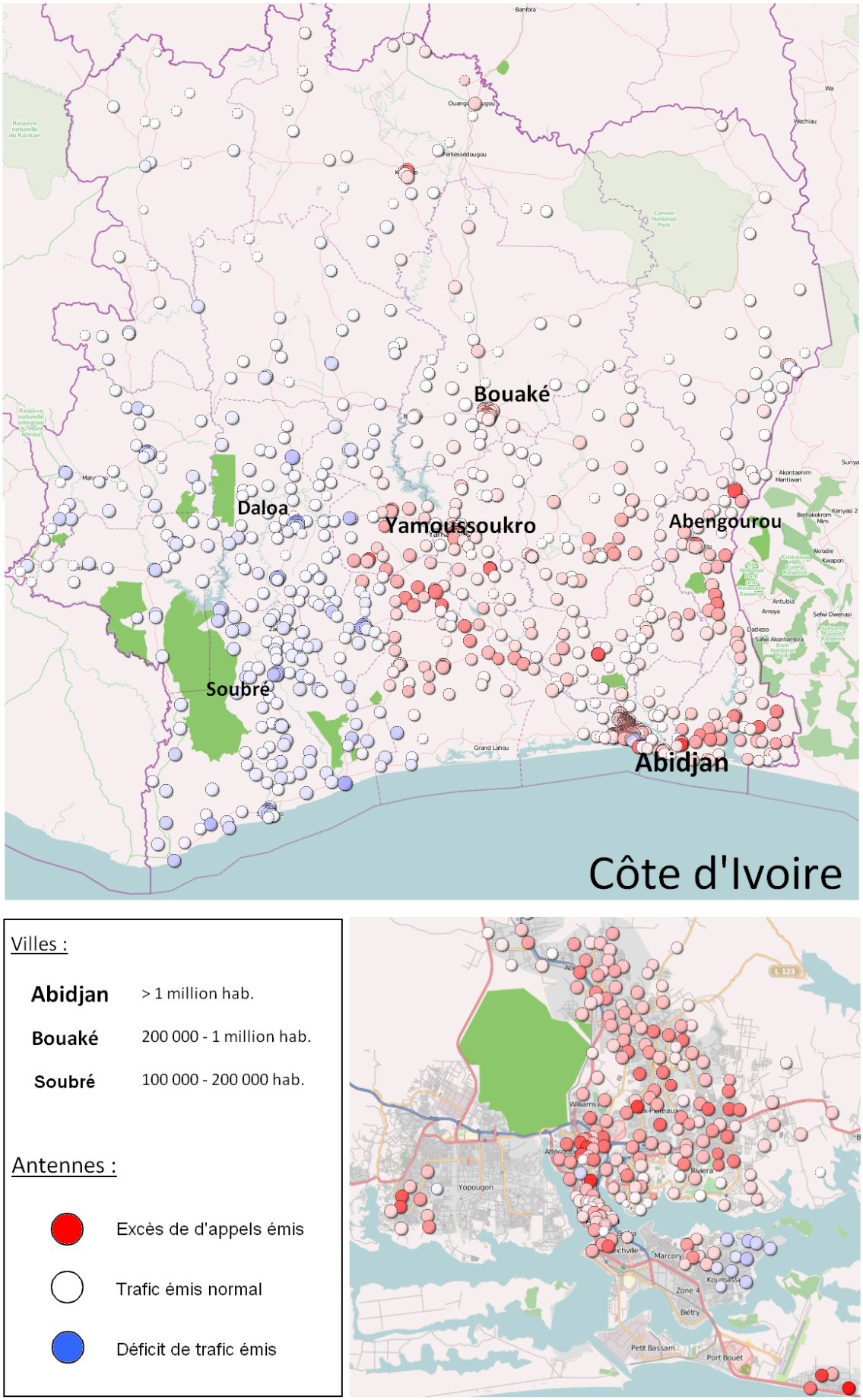


FIGURE 5.6 – Excès et déficits d'appels émis depuis les clusters d'antennes pour la période du 14 au 31 Décembre 2011.

période étudiée, c'est-à-dire la contribution à l'information mutuelle des couples clusters d'antennes / intervalles de temps. Le nombre d'intervalles de temps étant trop important nous décidons de le réduire à six, ce qui engendre une perte d'inertie inter-clusters de 30%. La partition des antennes est maintenue à son niveau optimal dans le but de préserver une information riche pour un ciblage local des excès et des déficits de trafic. Afin de déterminer la période la plus intéressante à étudier, la mesure d'intérêt est calculée sur les intervalles. Elle est maximale du 14 décembre 2011 au 31 Décembre 2011. Une cartographie est réalisée (voir figure 5.6), l'intervalle le plus intéressant est fixé et la contribution à l'information mutuelle entre l'intervalle étudié et chacun des clusters est calculée : s'il y a un excès (resp. un déficit) d'appels émis depuis un cluster entre le 14 décembre 2011 au 31 Décembre 2011 par rapport au trafic attendu, le cluster est coloré en rouge (resp. bleu) et sa couleur est d'autant plus intense que l'excès (resp. le déficit) est important.

Dans Guigourès *et al.* (2013), l'étude sur la France porte sur la période de Mai à Octobre 2007. En étudiant la discrétisation, on observe une alternance semaine / weekend hors vacances scolaires et un découpage en tranche de 7 jours pendant les vacances scolaires. De plus, la répartition des excès et des déficits d'appels sur le territoire en fonction des intervalles de temps montre une concentration des excès de trafic dans les villes en semaine, peu d'excès ou de déficits les weekend et des excès très importants de trafic sur les côtes en plein été.

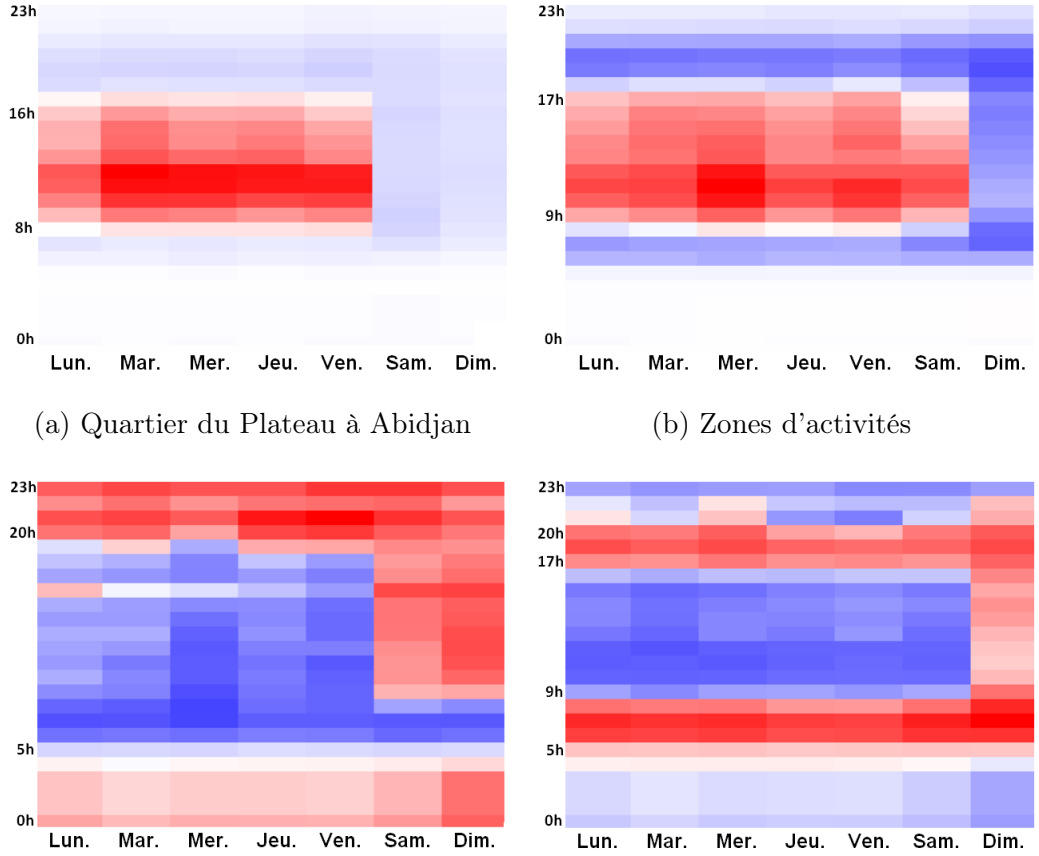
On n'observe aucun phénomène de ce type en Côte d'Ivoire. Tout l'Ouest du pays connaît des déficits d'appels du 14 décembre 2011 au 31 Décembre 2011, alors que l'Est connaît des excès. Ces excès sont locaux aux grandes villes comme Abidjan ou Bouaké. On les retrouve également dans les campagnes du Sud-Ouest de Yamoussoukro. Il est difficile de corréler ces résultats avec les événements dans le pays à cette période. Néanmoins, la zone Sud-Ouest du pays correspond à une région avec de fortes mobilités liées à l'activité agricole. La période des fêtes de fin d'année peut expliquer le déficit d'appels dans ces zones du pays par rapport au trafic observé les autres mois.

5.2.3 Étude des communications émises en fonction du jour de la semaine et de l'heure de la journée

Dans cette partie, nous cherchons à comprendre les habitudes des usagers en termes d'utilisation de leur téléphone mobile en fonction de l'heure de la journée, du jour de la semaine et de la zone où ils se trouvent. Nous proposons donc d'effectuer un triclustering des appels décrits par le jour de la semaine, l'heure de la journée et l'antenne émettrice de l'appel. Pour la même raison que précédemment, nous ignorons la destination de l'appel.

Au niveau le plus fin, nous obtenons un triclustering avec 806 clusters d'antennes émettrices, une discrétisation de la semaine en sept périodes et de l'heure de la journée en 22 intervalles. Une fois encore, il est nécessaire de réaliser

une simplification. Cependant, ce niveau de précision est acceptable pour la discrétisation de la semaine et de la journée. Nous simplifions donc la partition des antennes en quatre clusters, ce qui permet de conserver 51% de l'inertie inter-clusters de la partition. De manière à pouvoir interpréter les résultats, les clusters sont projetés sur une carte du pays (voir figure 5.8). D'autre part, pour chacun des clusters, nous traçons un calendrier hebdomadaire faisant apparaître les heures et jours de la semaine où on observe un excès ou un déficit d'appels par rapport au trafic attendu, émis depuis le cluster pendant la période étudiée (voir figure 5.7). Ces excès et déficit sont quantifiés par la contribution à l'information mutuelle entre le cluster d'antennes et le produit cartésien des intervalles d'heure et de semaines : $MI(X_1^M; X_2^M \times X_3^M)$ avec X_1^M la partition des antennes, X_2^M la discrétisation des jours de la semaine et X_3^M la discrétisation des heures de la journée.



(a) Quartier du Plateau à Abidjan

(b) Zones d'activités

(c) Zones résidentielles urbaines

(d) Zones rurales

FIGURE 5.7 – Calendrier des excès (en rouge) et déficits (en bleu) d'appels émis depuis les quatre clusters en fonction de l'heure et du jour de la semaine.

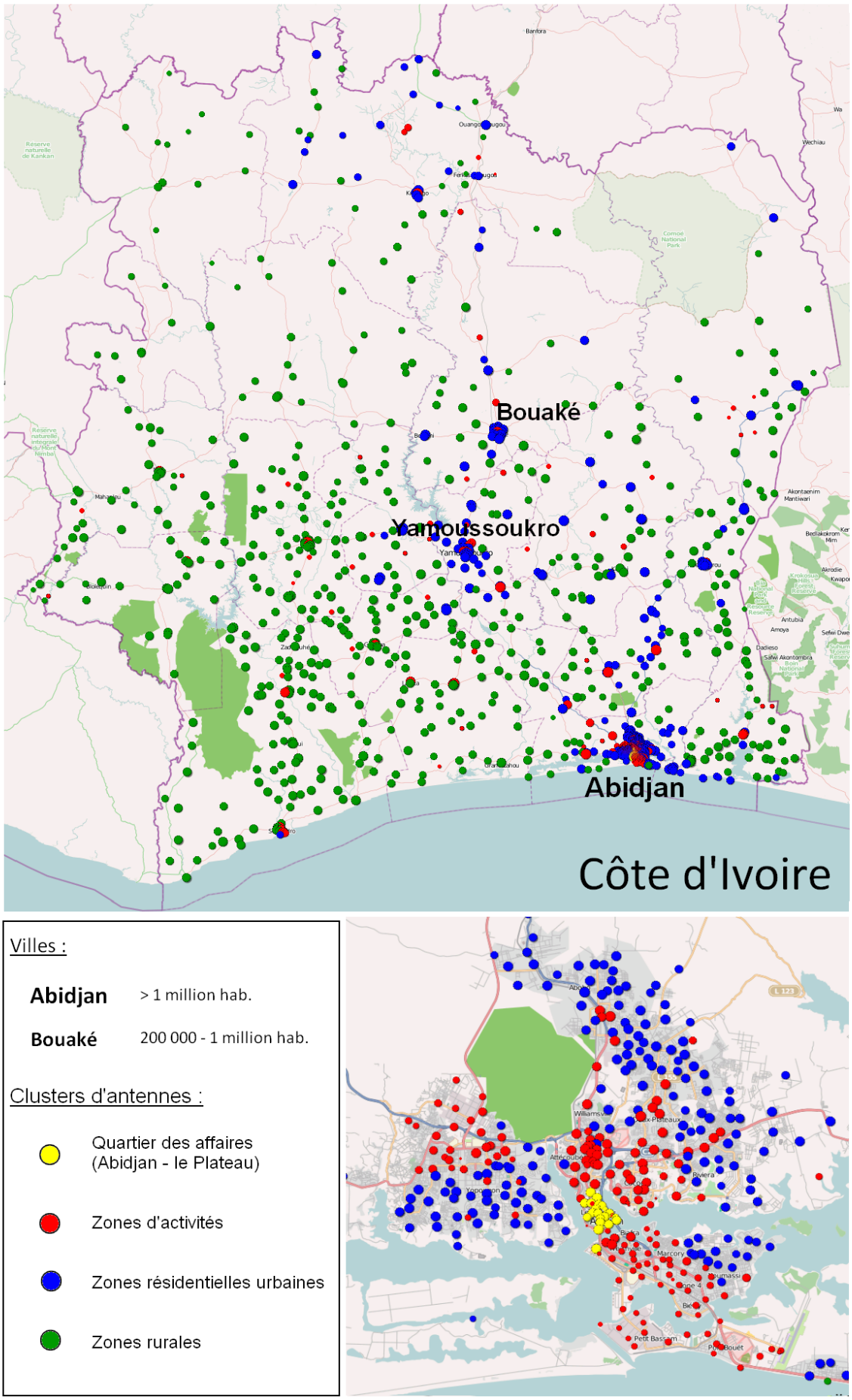


FIGURE 5.8 – Projection des quatre clusters sur une carte de la Côte d'Ivoire

Abidjan - Le Plateau. Le cluster d'antennes représentées par des points jaunes sur la carte d'Abidjan couvre exactement le quartier des affaires de la ville, qui est également le seul du pays. En observant le calendrier, il apparaît que l'activité téléphonique est en excès en semaine de 8-9h à 16-17h. Le reste du temps, le déficit de trafic y est léger. Cela signifie qu'aux horaires de bureau, la quantité d'appels est supérieure au trafic attendu et le reste du temps, elle est très légèrement inférieure. Ce comportement caractérise le type de quartier que couvre le cluster : un quartier d'affaire non résidentiel.

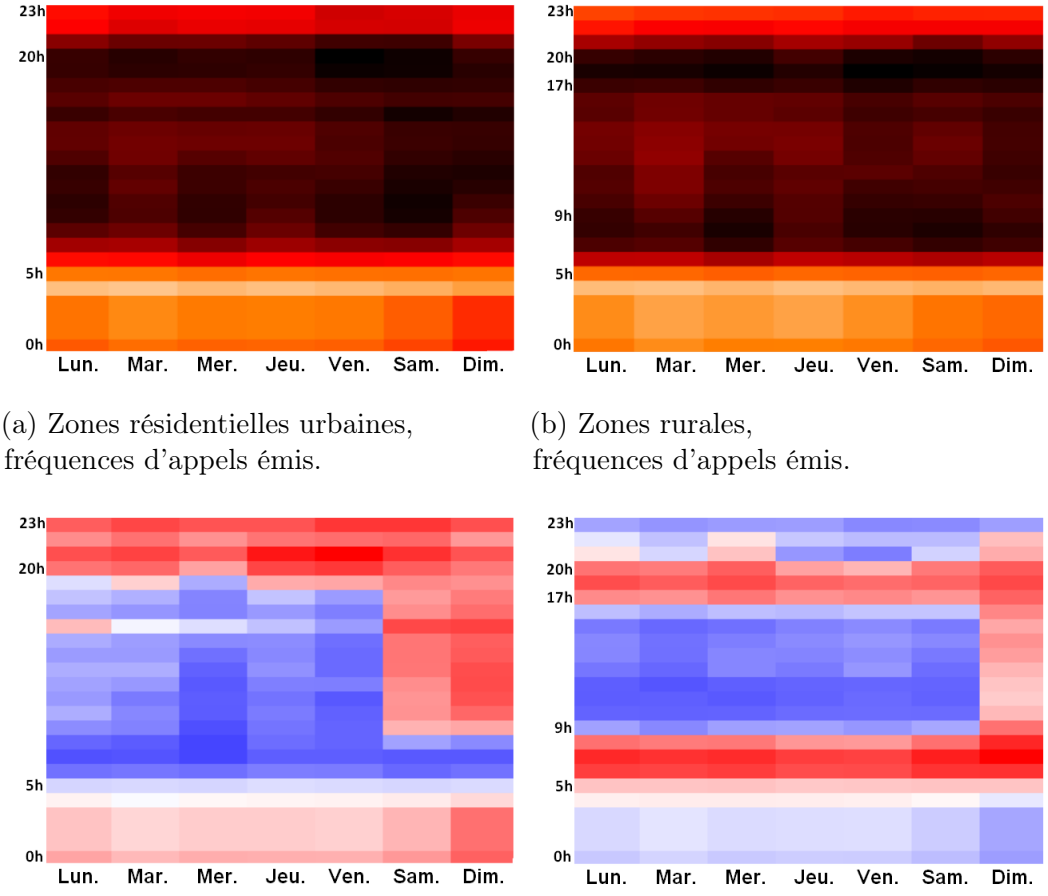
Les zones d'activité. Ces zones, modélisées par les points rouges sur la carte, sont dispersées sur l'ensemble du territoire à l'échelle nationale. Elles correspondent soit aux centres des villes, soit aux zones rurales avec une forte activité économique comme les plantations ou les mines. Au niveau d'Abidjan, on retrouve les zones industrielles de Marcory et du Nord de Yopougon, les quartiers commerçants de Treichville et d'Adjamé et l'Ouest de Cocody, le quartier des ambassades et de l'université. Ces zones sont caractérisées par des excès de trafic en semaine et le samedi de 9h à 17h, et d'important déficits d'appels à l'aube, le soir et le dimanche. On a donc une forte corrélation entre les périodes d'activités de ces quartiers et les usages téléphoniques.

Les zones résidentielles urbaines. Il s'agit des points bleus sur la carte. On retrouve les antennes avec ce profil au niveau des grandes villes comme Abidjan, Bouaké et Yamoussoukro, ainsi que dans le Nord. En s'intéressant à Abidjan, le cluster couvre les quartiers résidentiels de Youpougon, Abobo, Koumassi et Riviera (l'Est de Cocody). Il est à noter qu'avec un modèle plus fin, ce cluster se sépare en deux au niveau d'Abidjan séparant d'un côté les quartiers populaires d'Abobo, Koumassi et Yopougon et de l'autre le quartier aisé de Riviera. Les calendriers montrent une activité téléphonique le soir et le weekend en excès important, et en déficit pendant les périodes travaillées de la semaine, ce qui est caractéristique des zones résidentielles. On note des forts excès après et des déficits avant 20h, alors que les zones d'activités semblent réduire le trafic après $17h$. Ces excès d'appels nocturnes sont principalement liés aux offres tarifaires de l'opérateur Orange, plus avantageuses passée une certaine heure.

Les zones rurales. Représentées par des ronds verts, les antennes de ce cluster sont dispersées dans tout le pays et absentes d'Abidjan et des grandes villes en général. En regardant le calendrier, le profil de ces zones ressemble au profil des quartiers résidentiels, sauf que l'activité téléphonique se limite au matin entre 5h et 9h et au début de soirée entre 17h et 20h. Dans ces zones, seule la journée du dimanche connaît des excès de trafic, la semaine et le samedi connaissent un déficit d'appels dans la journée. Ce comportement s'explique par la nature de l'activité économique des campagnes : là où on observe des excès, les populations sont au village et dans les champs en journée. Notons

que l'usage du mobile dans les champs n'est pas moins fréquent que dans les villages. L'explication du déficit serait plutôt liée à une mauvaise couverture Orange dans les campagnes, les utilisateurs préférant utiliser un autre réseau.

Dans cette étude, le modèle de triclustering est très simplifié. À un niveau plus fin, une analyse plus complète pourrait être faite afin de caractériser les quartiers des villes ou des campagnes en fonction des habitudes des usagers. On pourrait envisager d'utiliser ces résultats afin d'optimiser les horaires d'ouverture et ainsi augmenter l'efficacité des points de vente en fonction du quartier.



(a) Zones résidentielles urbaines, fréquences d'appels émis.

(b) Zones rurales, fréquences d'appels émis.

(c) Zones résidentielles urbaines, contribution à l'information mutuelle.

(d) Zones rurales, contribution à l'information mutuelle.

FIGURE 5.9 – Calendrier des fréquences et des excès (resp. déficits) d'appels émis depuis deux des quatres clusters en fonction de l'heure et du jour de la semaine.

L'information mutuelle vs. la fréquence. Dans les précédents analyses, nous utilisons la contribution de chaque co-cluster à l'information mutuelle des

partitions du modèle pour modéliser les excès et déficits de trafic. Le nombre d'observations dans les co-clusters (c'est-à-dire la fréquence) est un concept plus simple à appréhender, cependant il ne permet pas de se comparer à la répartition moyenne des données sur les co-clusters. L'exemple présenté ici est très illustratif de l'intérêt d'utiliser la contribution à l'information mutuelle dans l'analyse exploratoire. En effet, le trafic mobile Ivoirien est principalement observé la journée et tous les jours de la semaine de manière similaire. Ainsi en analysant le nombre d'appels en fonction du jour de la semaine et de l'heure de la journée, localement à chaque cluster, il est impossible d'y discerner la moindre différence. La figure 5.9 présente un exemple particulièrement illustratif de l'intérêt d'étudier la contribution à l'information mutuelle plutôt que la fréquence des appels. Plus les cellules sont foncées, plus il y a d'appels au jour et à l'heure correspondants dans les deux calendriers 5.9.(a) et 5.9.(b). Dans les calendriers 5.9.(c) et 5.9.(d) permettent de mieux observer les différences entre les deux clusters grâce aux excès (en rouge) et aux déficits (en bleu) d'appels par rapport au trafic attendu aux périodes de la semaine et aux heures du calendrier.

5.3 Étude des communications entre mobiles et l'international

Les données dans cette partie sont des compte-rendus d'appels collectés en Janvier et Février 2012. Les données sont des communications entre le réseau mobile Ivoirien et l'international. Nous disposons des données suivantes :

- L'antenne ou le pays émetteur,
- L'antenne ou le pays récepteur,
- une estampille temporelle précise à la minute,
- Le type de communication (voix, SMS, etc).

Nous proposons trois études différentes. La première consiste en l'étude des communications passées depuis les antennes mobiles Ivoiriennes vers l'international. La seconde est une étude du trafic émis depuis l'international vers les mobiles Ivoiriens en fonction de l'heure. Enfin, dans la dernière étude, nous proposons une analyse des communications passées depuis l'international vers la Côte d'Ivoire, en fonction de l'heure, du jour de la semaine et du type de communication.

5.3.1 Analyse du trafic entre les antennes Ivoiriennes et l'international

Dans cette étude, nous nous intéressons aux communications mobiles passées depuis le réseau Ivoirien vers l'international. Nous disposons d'un total de 22 063 409 communications de différentes natures comme la voix, les messages ou l'usage de l'internet mobile. Nos données sont donc modélisées par un graphe

biparti dont l'un des ensembles de nœuds correspond aux antennes Ivoiriennes et l'autre aux pays.

Les résultats initiaux retrouvés sont très fins et nécessitent une simplification : nous obtenons 794 clusters d'antennes et 65 clusters de pays. En réduisant le nombre de clusters d'antennes à 15 et le nombre de clusters de pays à 10, nous maintenons 80% de l'information du modèle. Les 10 clusters trouvés sont listés dans la table 5.4.

Id Cluster	Intérêt	Trafic émis	Nombre de pays
Burkina Faso	1	34,07%	1
Cluster 1[1]	0,2896	16,24%	612
Cluster 2[1]	0,1399	18,43%	1
Cluster 3[1]	0,0951	2,16%	2
Cluster 4[1]	0,0767	5,08%	2
Cluster 5[1]	0,0537	9,05%	1
Cluster 6[1]	0,0358	6,78%	146
Cluster 7[1]	0,0342	4,54%	13
Cluster 8[1]	0,0311	1,52%	1
Cluster 9[1]	0,0137	2,11%	1

TABLE 5.4 – Dix clusters de pays. L'identifiant du cluster est choisi tel que le pays soit le plus représentatif du cluster. L'intérêt du cluster et le trafic émis sont également indiqués.

Pour construire la table 5.4, nous avons utilisé la démarche méthodologique introduite dans la section 5.1.2 : nous avons calculé la typicité des valeurs (ici les pays) des dix clusters et utilisé la valeur la plus typique comme étiquette du cluster. Lorsque deux pays ont des typicités proches, nous utilisons les deux noms pour étiqueter le cluster. La mesure d'intérêt est présentée afin que l'on puisse se focaliser sur les clusters les plus structurants de la partition des pays, donc ceux apportant le plus d'information. Les clusters sont triés par intérêt. Nous regardons également le trafic émis depuis les clusters afin de compléter l'analyse de l'intérêt des clusters.

La table 5.4 montre un déséquilibre des clusters : le trafic émis n'est pas réparti sur l'ensemble des clusters. On peut noter certains clusters de pays qui interagissent énormément avec les mobiles Ivoiriens comme le Burkina Faso ou le cluster 1. D'autres comme le cluster 3 ou le cluster 8 reçoivent peu de communications depuis la Côte d'Ivoire mais ces dernières sont suffisamment caractéristiques pour que ces pays constituent des clusters.

Nous nous focalisons sur le cluster les plus intéressants : le Burkina Faso. Ce cluster maximise la mesure d'intérêt et contient seulement le Burkina Faso. On s'intéresse à la répartition des communications provenant des antennes de Côte

1. Pour des raisons de confidentialité, nous ne donnons pas de détails sur ces clusters.

d'Ivoire vers ce cluster. On analyse pour cela la contribution à l'information mutuelle du couple cluster d'antennes / cluster du Burkina Faso. Cela nous permet de localiser les zones depuis lesquelles les communications sont en excès (ou en déficits) vers le Burkina Faso par rapport au trafic attendu. La carte de la figure 5.10 permet de visualiser les régions depuis lesquelles on observe ces excès et déficits.

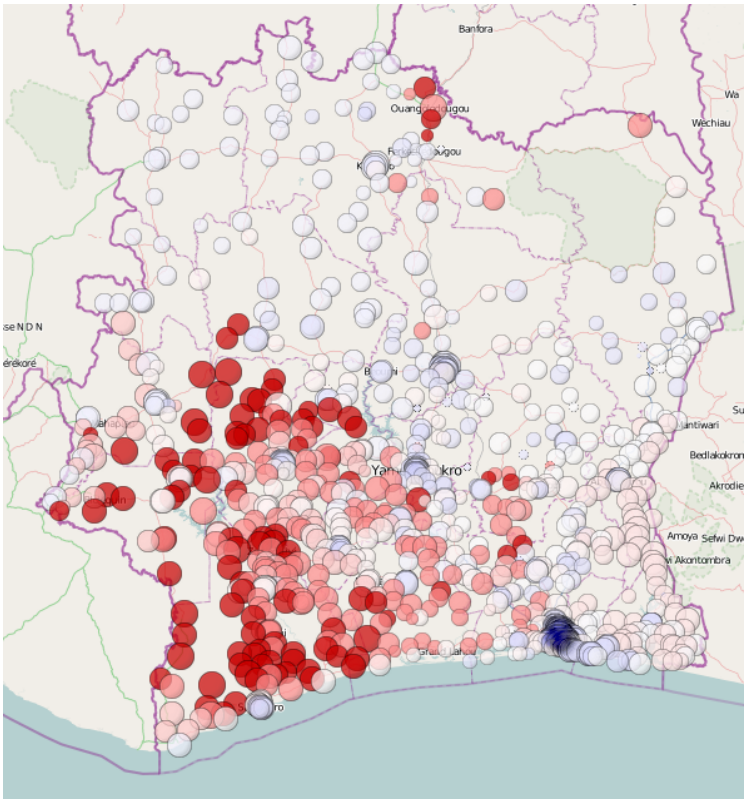


FIGURE 5.10 – Excès (en rouge) et déficits (en bleu) de communications vers le Burkina Faso.

On voit apparaître deux zones depuis lesquelles les communications sont émises en excès. La première se trouve au Nord du pays près de la frontière avec le Burkina Faso. On a donc affaire à des communications transfrontalières, ces excès s'expliquent donc par la proximité géographique du cluster d'antennes et du pays représentatif de son cluster. L'autre zone est située au Sud-Ouest du pays. Cette partie de la Côte d'Ivoire a une activité principalement agricole. Les cultures du cacao, du café ou encore de l'hévéa y sont intensives. Les migrations des Burkinabés vers ces zones du pays sont fréquentes aux saisons des récoltes, ce qui explique qu'on y observe ces excès de trafics. Notons les déficits de communications émises depuis les principales villes du pays que sont Yamoussoukro, Abidjan et Bouaké, ainsi que les capitales régionales comme Man et San Pédro.

Cette étude nous a permis de mettre en évidence les zones dans lesquelles on observait des excès et déficits de trafics vers certains pays. Dans un deuxième

temps, nous proposons d'étudier le trafic dans le sens inverse, c'est-à-dire depuis les pays vers les mobiles Ivoiriens. La première constatation est une asymétrie du trafic, le nombre de communications de l'international vers les mobiles Ivoiriens est presque deux fois moins important que le trafic dans le sens inverse. Une étude des clusters de pays obtenus montre une partition différente des pays. Cependant l'étude de la répartition des appels sur le territoire pour le Burkina Faso est similaire dans les deux sens du trafic.

5.3.2 Analyse du trafic émis depuis l'international vers les antennes Ivoiriennes en fonction de l'heure

Dans cette étude, nous nous intéressons aux communications mobiles passées depuis le réseau Ivoirien vers l'international en fonction de l'heure de la journée. Nous disposons toujours d'un total de 22 063 409 communications. Nos données sont modélisées par un graphe biparti temporel dont l'un des ensembles de nœuds correspond aux antennes Ivoiriennes et l'autre aux pays et les arcs entre les deux sont estampillés de l'heure de la communication.

Les résultats initiaux retrouvés sont très fins et nécessitent une simplification : nous obtenons 286 clusters d'antennes, 33 clusters de pays et 10 plages horaires. En réduisant simultanément le nombre de clusters d'antennes à 12, le nombre de clusters de pays à 11 et le nombre d'intervalles de temps à 6, nous maintenons 80% de l'information du modèle.

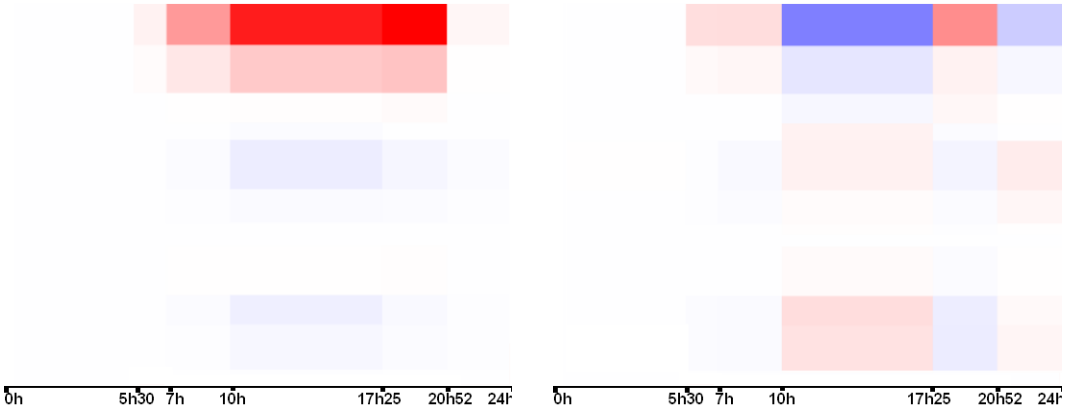
Nous allons, dans cette étude, chercher à comprendre comment sont structurés les appels provenant de l'international en fonction des zones couvertes par les clusters d'antennes réceptrices et de l'heure de la journée. Comme précédemment, nous étudions les excès et les déficits d'appels.

L'importance du choix de la visualisation Jusqu'à présent, dans le cas d'un co-clustering à plus de deux dimensions, nous avons utilisé la fonction de contraste afin de visualiser les excès et les déficits de trafic. Cela nous a servi, lors de la construction des calendriers semainiers, à comprendre de quels clusters provenaient les excès et déficits d'appels en fonction du jour de la semaine et de l'heure de la journée. On a donc utilisé le contraste défini comme une information mutuelle entre la partition des antennes et le produit cartésien des discrétisations des heures et des jours de la semaine $MI(X_1^M, X_2^M \times X_3^M)$, avec X_1^M la partition des antennes, X_2^M la discrétisation des jours de la semaine et X_3^M la discrétisation des heures.

Nous cherchons à caractériser le comportement des utilisateurs, appelant certains groupes de pays en fonction de leur position géographique (modélisée par l'antenne utilisée) et l'heure de la journée. Tracer le contraste permet de voir quel groupe de pays est appelé en excès depuis un groupe d'antennes à une plage horaire donnée. Le problème est la forte corrélation entre clusters d'antennes et de pays, montrée dans la section 5.3.1 pour certains pays comme le Burkina Faso. Ainsi, tracer le contraste ne ferait que mettre en évidence ces

corrélations entre le pays émetteur des communications et le cluster d'antennes qui les reçoit, plus ou moins important en fonction de l'intensité du trafic de la plage horaires (voir figure 5.11.(a)).

Dans cette étude, on préfère utiliser l'information mutuelle conditionnelle au cluster de pays étudié. Il s'agit de mesurer les excès et déficits de communications reçues par les groupes d'antennes aux différents intervalles de temps, conditionnellement à un pays émetteur. On trace donc $MI(X_2^M, X_3^M|X_1^M)$ avec X_1^M la partition des pays, X_2^M la partition des antennes et X_3^M la discrétisation de l'heure. La figure 5.11 est une illustration des résultats obtenus sur le cluster du Burkina Faso. En affichant le contraste (figure 5.11.(a)), on obtient des excès de trafic sur deux clusters d'antennes. Et ces excès se répartissent de la même manière que la fréquence sur les heures de la journée : une forte intensité en journée, faible le soir et le matin et quasi nulle la nuit. L'information mutuelle conditionnelle au cluster du Burkina Faso (figure 5.11.(b)) nous permet de mettre en évidence un déficit de communications important en journée principalement vers un cluster de la Côte d'Ivoire par rapport au trafic habituel vers ce cluster et par rapport au trafic émis depuis le Burkina à cette période de la journée vers l'ensemble des antennes de Côte d'Ivoire. Ce déficit n'apparaît pas lorsque le contraste est étudié.

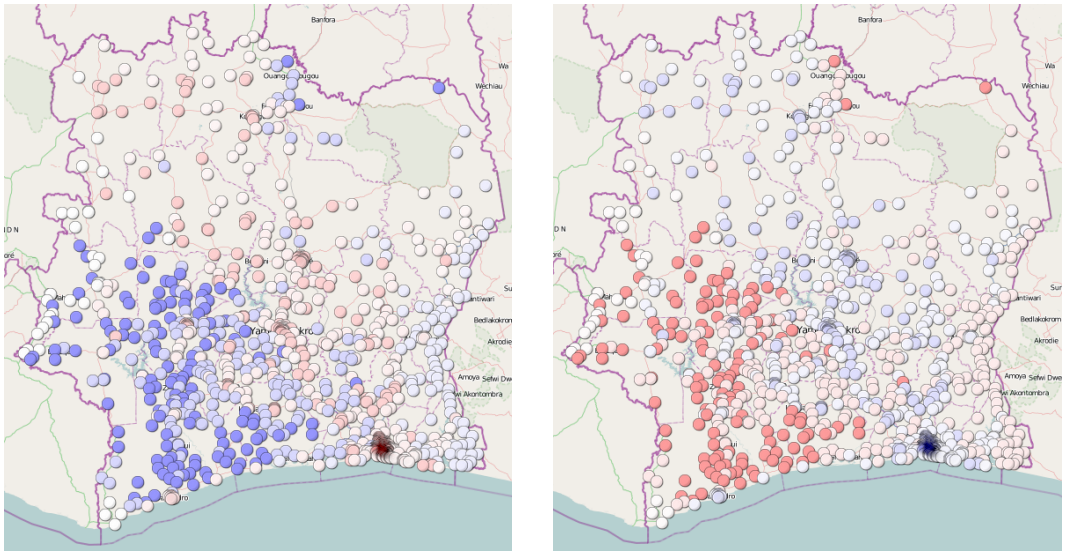


(a) Contraste (b) Information mutuelle conditionnelle

FIGURE 5.11 – Contraste et information mutuelle conditionnelle au cluster du Burkina Faso. En ligne, les clusters d'antennes et en colonne les plages horaires.

On s'intéresse au cluster de pays le plus intéressant : le Burkina Faso. Ce cluster se limite à un pays. C'est de loin le pays qui émet le plus de communications vers les mobiles de Côte d'Ivoire et également celui qui maximise la mesure d'intérêt de la partition des pays. La carte de la contribution à l'information mutuelle apparaît dans la figure 5.12.

Comme on l'a vu dans une précédente étude, c'est dans le Sud-Ouest du pays que le trafic est le plus important, ce qui explique les forts contrastes dans les excès et déficits de communications observés dans la figure 5.12. En journée, le déficit est important dans cette zone et le soir, les communications



(a) $[10h00; 17h25[$ (b) $[17h25; 20h52[$

FIGURE 5.12 – Information mutuelles entre les clusters d'antennes réceptrices Ivoiriennes et les intervalles de temps $[10h00; 17h25[$ et $[17h25; 20h52[$, conditionnellement au cluster du Burkina Faso.

sont en excès ce qui est caractéristique du comportement observé dans les zones rurales dans l'étude de la section 5.3.1. Cela nous conforte dans l'idée que les populations venant du Burkina Faso occupent principalement des emplois agricoles.

5.3.3 Analyse du trafic émis depuis l'international vers les mobiles Ivoiriens en fonction de l'heure et du type de service

Dans cette étude, nous étudions toutes les variables descriptives des données : le pays émetteur, l'antenne réceptrice, le jour de la semaine, l'heure de la journée et le type de service (SMS, appel...). Nous disposons toujours d'un total de 22 063 409 communications. On réalise donc un pentaclustering (co-clustering en cinq dimensions).

L'approche MODL ne discrétise pas les jours de la semaine. La variable est donc éliminée du modèle car elle n'est pas discriminante. On a donc une grille de co-clustering avec quatre dimensions : 186 clusters d'antennes, 32 clusters de pays, 8 intervalles de temps et 2 clusters de services. En réduisant le nombre de clusters d'antennes à sept, le nombre de clusters de pays à sept et le nombre d'intervalles de temps à cinq, nous maintenons 80% de l'information du modèle. Nous conservons les deux clusters de services : l'un ne contient que les SMS et l'autre, les autres services dont 99,98% concernent les appels vocaux.

Le type de service est très structurant pour les données et fait apparaître de nouveaux clusters de pays. Le cluster du Burkina Faso ne contient toujours qu'un pays. Les usagers ne passent pratiquement que des appels et très peu de SMS. Ainsi, nous obtenons des résultats similaires à la précédente étude, à savoir des excès d'appels le matin et en soirée, et un trafic normal sinon.

5.4 Étude de mobilité

La troisième base de données étudiée correspond à un relevé des positions de 50 000 usagers choisis aléatoirement pendant deux semaines. Un usager est localisé par l'antenne utilisée lorsqu'il utilise le réseau. Cette base de données est composée de 55 millions d'enregistrements sur 1 214 antennes du réseau. On dispose également d'une estampille temporelle précise à la minute. La première analyse est une étude des trajectoires utilisateurs. Une trajectoire est définie comme un parcours de deux semaines d'un utilisateur dans le réseau d'antennes. Dans une seconde étude, on cherche à détecter les usagers avec un profil « voyageur » par une étude des courbes d'utilisation des antennes pour chaque utilisateur. Enfin, nous proposons une analyse des déplacements des « voyageurs » en fonction du jour de la semaine et de l'heure de la journée.

5.4.1 Étude des trajectoires

On s'intéresse dans un premier temps aux trajectoires qui peuvent se modéliser comme un graphe biparti avec 50 000 nœuds « usagers », 1 214 nœuds « antennes » et 55 millions d'arcs. On cherche donc deux partitions : une partition des usagers et une partition des antennes. Ainsi les utilisateurs qui utilisent les mêmes antennes, et les antennes couvrant les mêmes utilisateurs sont groupés.

	Partitions des usagers	**Partitions des antennes**
Nombre de clusters	6686	606
Taux de dispersion des clusters	53,76%	36,20%

TABLE 5.5 – Caractéristiques des partitions

Le tableau 5.5 donne les caractéristiques relatives aux segmentations des utilisateurs et des antennes. Le nombre de clusters est très important. Dans le cas des antennes, on obtient 606 clusters, ce qui est très fin (environ deux antennes par clusters en moyenne), mais bien moins que dans l'étude du trafic inter-antennes. Dans le cas des utilisateurs, on a un grand nombre de clusters contenant un seul individu. Ceci est du à la richesse des données : on dispose de plus de mille points de mesures en moyenne par usager. La valeur du taux de dispersion des clusters d'antennes (inertie inter-clusters normalisée) est de

36%, ce qui est important mais inférieur à la dispersion obtenue par étude du trafic inter-antennes. Cela signifie que la corrélation entre les partitions des antennes émettrices et réceptrices est plus importante que la corrélation entre les partitions des usagers et des antennes qu'ils utilisent. Cependant, la partition des antennes est une excellente explication de la segmentation des clients : on a un taux de dispersion de la partition de près de 54%. Cette forte inertie ainsi que la finesse des clusters obtenus montre que la répartition des appels dans les clusters d'antennes peut permettre d'identifier la plupart des 50 000 individus de la base de données.

En s'intéressant aux biclusters, on observe que les utilisateurs du réseau se déplacent très peu : leur activité se distribue principalement sur un seul cluster d'antennes. Si on utilise les variables temporelles dans cette étude, on obtient aucune discrétisation car la corrélation entre les segmentations des antennes et des utilisateurs est bien plus informative, à l'image de ce qui est observé dans l'étude du trafic inter-antennes. Cependant, dans la précédente étude, le graphe d'appels était biparti avec deux ensembles d'antennes identiques. Les partitions des antennes émettrices et réceptrices étaient identiques, ce qui nous a permis de supprimer l'un des deux ensembles d'antennes des autres études. Ici, nous ne pouvons nous contenter de supprimer une variable. On filtre donc les données de manière à ne conserver que les usagers caractérisés comme « voyageurs ».

5.4.2 Étude des courbes de trafics par utilisateur

À partir des trajectoires, nous avons créé un nouveau jeu de données afin de filtrer les utilisateurs statiques. Les nouvelles données sont décrites par trois variables :

- *l'identifiant utilisateur* : variable nominale issue des données de mobilités,
- *rang d'utilisation des antennes* : les antennes sont classées en fonction de leur fréquence d'utilisation par l'utilisateur, la variable est continue,
- *fréquence d'utilisation de l'antenne* : nombre d'utilisations de l'antenne par l'usager, variable continue.

On se place donc dans le cadre du clustering de courbes décrit dans la section 3.4. Les courbes sont au nombre de 50 000, soit le nombre total d'utilisateurs. Le nombre de points total est d'environ 2,8 millions, soit en moyenne 56 par courbe, ce qui correspond au nombre moyen d'antennes utilisé par chaque usager sur la période d'observation.

L'application d'un triclustering de deux variables continues et une nominale nous permet de grouper les utilisateurs en 128 clusters et de discrétiser les rangs en 28 intervalles et les fréquences en 16 intervalles. Ce niveau est trop fin pour filtrer les données. En conservant 60% du taux d'information du modèle, nous avons trois profils utilisateurs, sept intervalles de rangs d'antennes et cinq intervalles de fréquences d'appels.

La figure 5.13 présente la courbe moyenne de la fréquence d'appels en fonction du rang d'utilisation de l'antenne pour l'ensemble des données. Cette

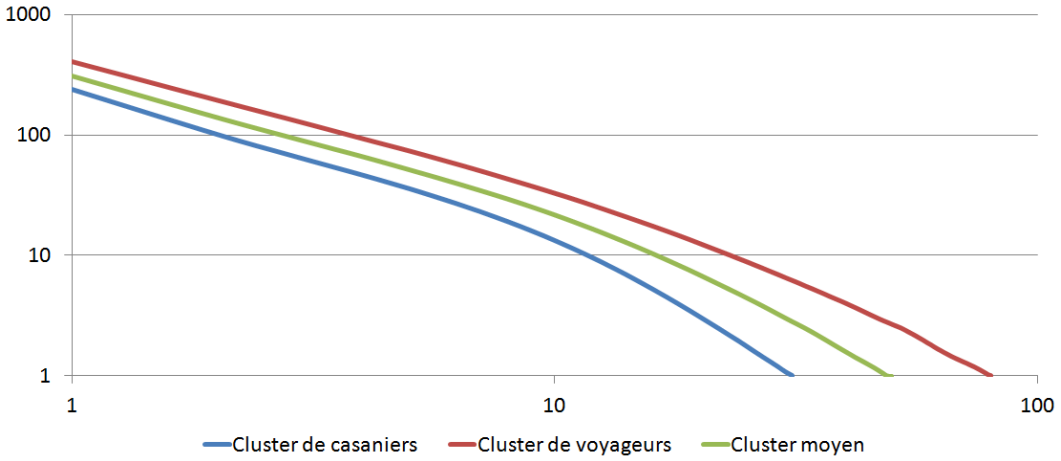


FIGURE 5.13 – Courbes moyennes de la fréquence d'appels en fonction du rang de l'antenne pour les trois clusters. Échelle logarithmique.

courbe nous permet de conforter l'idée que les utilisateurs sont globalement peu mobiles : en moyenne sur deux semaines, les utilisateurs se connectent près de 300 fois (26% des connexions) sur l'antenne qu'ils utilisent le plus et environ 80 fois (7,6% des connexions) en dehors de leurs vingt antennes les plus utilisées. Nous avons également tracé les courbes moyennes par cluster. Les trois courbes sont très similaires à la courbe moyenne des données. L'analyse de ces courbes permet de caractériser les clusters qu'elles représentent. Un cluster d'utilisateurs « casaniers » est représenté par la courbe prenant les valeurs de fréquence en-dessous de la courbe moyenne. Ces utilisateurs sont caractérisés par l'utilisation de peu d'antennes différentes. Au contraire, la courbe des utilisateurs dits « voyageurs » présente des fréquences au-dessus de la moyenne des données, ce qui signifie que ces usagers utilisent un nombre plus important d'antennes du réseau et donc présentent une plus grande mobilité. Enfin, un dernier cluster a une courbe de trafic très proche de la courbe moyenne et est donc désigné comme cluster d'utilisateurs à mobilité moyenne. Notons pour ce cluster que la mesure d'intérêt est faible, ce qui nous montre que la répartition des communications de ce groupe est proche de la répartition moyenne des données.

Notons que les courbes d'utilisation des antennes par les utilisateurs suivent une loi de puissance. Ainsi on aurait pu directement chercher les paramètres de cette loi et appliquer un simple algorithme de clustering afin de regrouper les courbes.

Le cluster qui nous intéresse est celui qui groupe les utilisateurs les plus mobiles, c'est-à-dire le groupe d'utilisateurs dits « voyageurs ». Afin de mieux comprendre le comportement de ces utilisateurs, nous proposons d'étudier l'information mutuelle entre le groupe d'utilisateurs mobiles et les biclusters formés par les discrétisation du rang des antennes et des fréquences d'appels, c'est-à-dire : $MI(X_1^M; X_2^M \times X_3^M)$ avec X_1^M la partition des utilisateurs, X_2^M

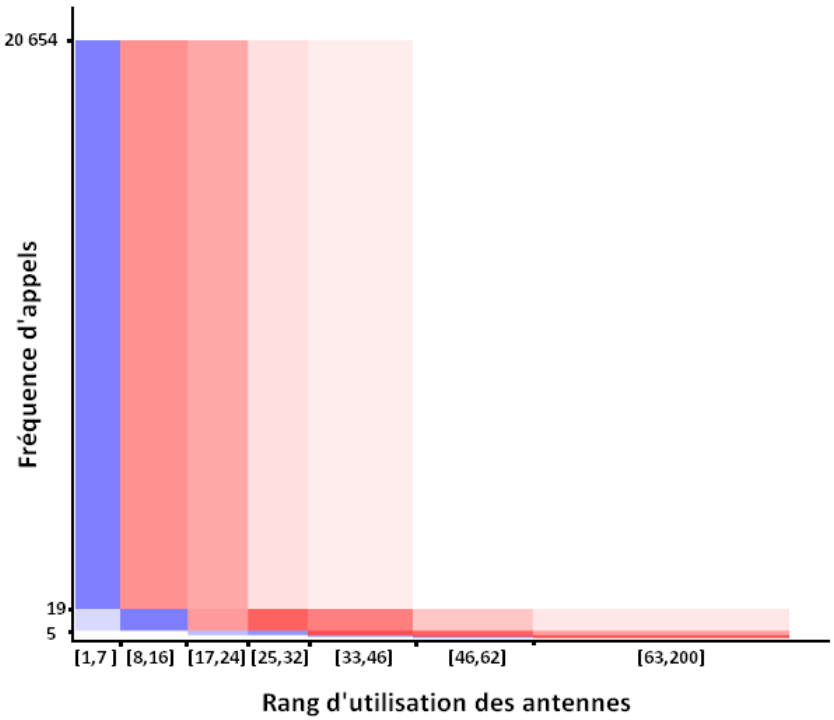


FIGURE 5.14 – Information mutuelle entre le groupe d'utilisateurs dits « Voyageurs » et les biclusters formés par les discrétisations du rang des antennes et des fréquences d'appels.

la discrétisation du rang des antennes et X_3^M la discrétisation de la fréquence d'utilisation des antennes. La figure 5.14 présente ces résultats. Les cases bleues montrent une sous-représentativité des usagers du cluster utilisant en majorité entre une et sept antennes. Au contraire, les cases rouges montrent une sur-représentativité des utilisateurs utilisant significativement plus de huit antennes. Ces utilisateurs sont donc plus mobiles et vont nous servir à construire une base de données d'utilisateurs mobiles nous permettant d'exploiter l'estampille temporelle des trajectoires.

5.4.3 Étude des trajectoires en fonction du jour de la semaine et de l'heure de la journée

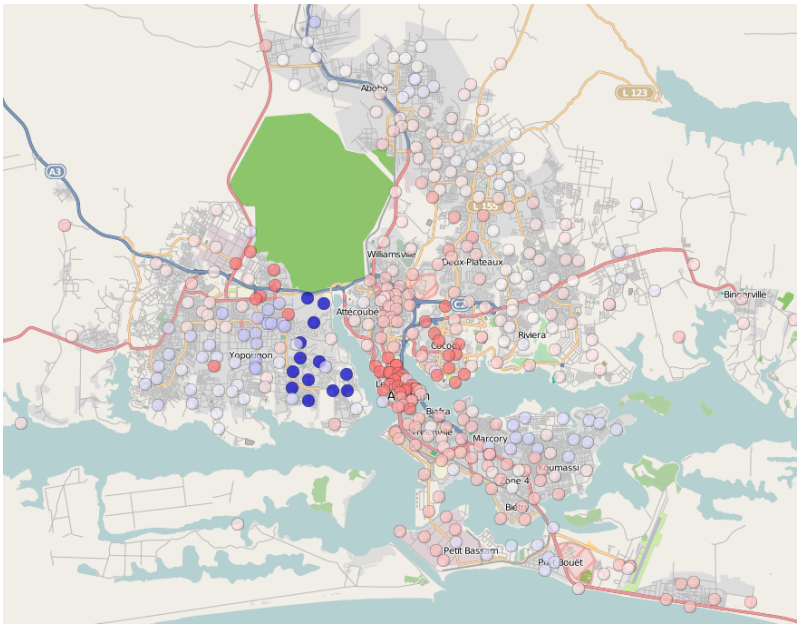
Une fois la base filtrée afin de ne conserver que les utilisateurs les plus mobiles, on se retrouve avec des données avec une volumétrie moins importante : 6 894 utilisateurs différents, 1 214 antennes parcourues et 12 345 601 utilisations du réseau collectées. On propose dans cette étude de segmenter les utilisateurs en fonction de leurs usages en termes de téléphonie : les utilisateurs sont groupés s'ils utilisent les antennes d'un même cluster, aux mêmes heures de la journée et les mêmes jours de la semaine. On réalise donc un tétraclustering (co-clustering

en quatre dimensions) entre les variables utilisateurs, antennes, jour de la semaine et heure de la journée.

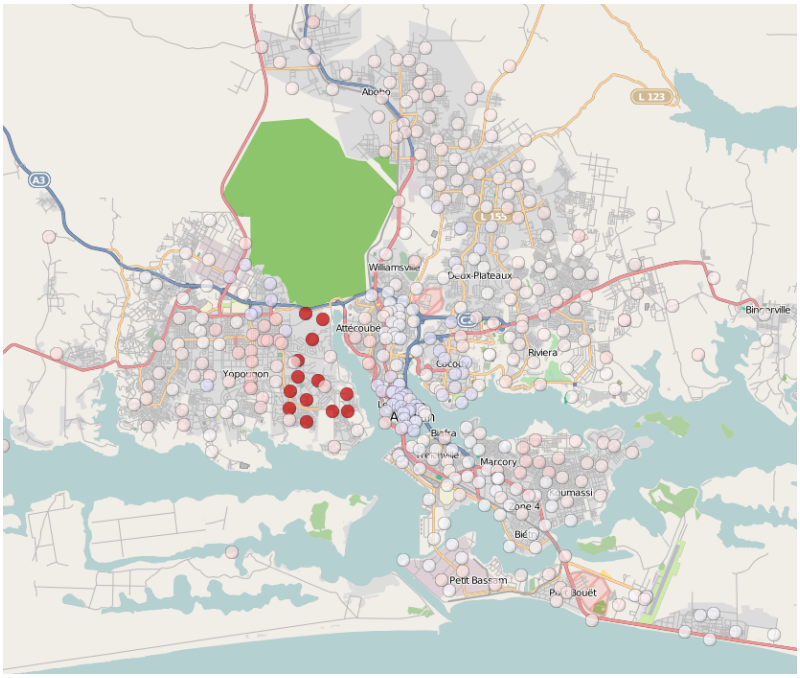
Au niveau le plus fin, nous obtenons 237 groupes d'utilisateurs, 218 clusters d'antennes, 7 découpages de la semaine et 3 intervalles d'heures. On cherche, via cette étude, à caractériser le comportement des utilisateurs en termes de mobilité. On choisit donc un exemple illustratif en se focalisant sur un groupe d'utilisateurs. Ce groupe est caractérisé par une forte valeur d'intérêt et donc un comportement qui se détache du comportement moyen des utilisateurs étudiés. Nous étudions, pour ce groupe, les excès et les déficits d'utilisations d'antennes. Afin d'avoir des groupes d'utilisateurs plus peuplés, une discrétisation des variables temporelles moins précise et une segmentation géographique moins fine, on simplifie le modèle de manière à conserver 50% de l'information du modèle le plus fin. Cela revient à conserver 40 clusters d'utilisateurs et d'antennes et 2 intervalles de semaine et d'heure. La semaine est donc divisée en deux parties : les jours ouvrés et le weekend. Quant à la coupure de l'heure, elle a lieu à 18h. Notons qu'on a deux intervalles : $[0h, 18h[$ et $[18h, 24h[$. La coupure de minuit est artificielle et donc non pertinente. La coupure de $18h$ est la dernière coupure réalisée dans la hiérarchie des intervalles d'heures, on aurait donc pu envisager de faire démarrer la journée à $18h$, afin d'avoir une coupure plus pertinente que minuit. Cependant, nous faisons le choix de la conserver telle qu'elle afin d'avoir une interprétation plus aisée de la discrétisation : il est plus facile d'étudier un découpage horaire sur la période $0h$ - $24h$, plutôt que sur la période $18h$ - $42h$.

Nous proposons de tracer les cartes d'utilisation des antennes pour deux périodes : la semaine avant $18h$ et la semaine après $18h$. L'étude du weekend montre un comportement similaire à la semaine en soirée. Les cartes sont présentées dans la figure 5.15.

La figure 5.15 nous permet de faire une analyse du comportement des utilisateurs du cluster sélectionné. On voit clairement apparaître une zone très contrastée : les antennes de la partie Est de Yopougon sont utilisées en fort excès le soir et le weekend alors qu'elles connaissent un fort déficit d'appels en semaine avant $18h$. On peut donc supposer que les utilisateurs résident tous dans ce quartier. D'autre part cette hypothèse est confortée par la nature de la zone couverte par ce cluster d'antennes : il s'agit d'une zone résidentielle. Les autres zones sont beaucoup moins contrastées. On peut noter que les excès de trafic en journée ont lieu majoritairement autour de trois zones : le Plateau, Cocody-Ambassades et la Zone industrielle de Yopougon. Le plateau est le cœur économique du pays, beaucoup d'institutions et d'entreprises y sont basées. Cocody-Ambassades est, comme son nom l'indique, le quartier des ambassades mais également celui d'un des principaux hôpitaux de la ville et de l'université (bien que celle-ci fût fermée à l'époque de la collecte des données). Enfin, le quartier de la zone industrielle de Yopougon est une zone d'activité à caractère industriel où travaillent plusieurs milliers d'ouvriers. Le point commun entre toutes ces zones est leur activité économique. On peut donc supposer que les utilisateurs du cluster étudié travaillent dans ces endroits.



(a) La semaine en journée



(b) La semaine en soirée

FIGURE 5.15 – Pour un groupe d'utilisateur sélectionné, excès et déficits d'utilisation des antennes mobiles. Focus sur Abidjan.

Les utilisateurs ont été regroupés à cause de leur lieu d'habitation plus que leur lieu de travail. On a donc une zone très délimitée où le trafic est très important de la part de ce groupe, tandis que les zones où les utilisateurs

travaillent sont disséminées dans Abidjan. C'est la raison qui fait que les excès et déficits sont plus contrastés au niveau de lieux de résidence des utilisateurs.

Pour résumer, on a pu ici caractériser un groupe d'usagers du réseau et comprendre leur utilisation du réseau grâce à l'analyse de leur mobilité. D'autre part, par simple étude de leurs heures et antennes de connexions et avec un peu de connaissance des quartiers de la ville, on a pu dresser des hypothèses quant à la nature d'activité : on a affaire à des actifs travaillant dans trois zones bien distinctes et habitant dans la partie Est du quartier résidentiel de Yopougon.

5.5 Conclusion de l'étude

L'objectif de cette dernière partie a été d'illustrer l'utilisation du co-clustering pour extraire des informations différentes d'une base de données à partir des modélisations proposées dans le chapitre 3. Nous montrons également l'intérêt des outils d'analyse exploratoire introduits dans le chapitre 4 et proposons une méthodologie d'analyse.

Analyse du trafic mobile entre antennes Les données sont les appels mobiles de la Côte d'Ivoire qui sont décrits par l'antenne émettrice, l'antenne réceptrice et une estampille temporelle. À partir des ces données, nous avons proposé une étude du graphe de communications entre-antennes, une étude du trafic émis à partir des antennes en fonction du jour de l'année et enfin une étude du trafic émis en fonction du jour de la semaine et de l'heure de la journée. Toutes ces études nous ont permis de caractériser le territoire Ivoirien en fonction des habitudes d'appels des usagers de téléphones mobiles. On a pu mettre en évidence une tendance des utilisateurs à appeler dans des zones géographiques très délimitées, dont les antennes les plus caractéristiques de la zones se trouvent en général dans des villes importantes avec un rayonnement national ou régional. L'étude avec la variable date nous a permis de mettre en évidence des périodes où les données sont manquantes en groupant des antennes ayant connu des pannes aux mêmes périodes, ce qui nous a permis de filtrer les données afin de recommencer une étude plus fiable. Ceci illustre la démarche itérative du processus de data mining introduit dans l'introduction. Enfin l'analyse du trafic en fonction du jour de la semaine et de l'heure de la journée nous a révélé une segmentation géographique, différente de celle obtenue dans l'étude du graphe de communications, et corrélée avec le profil socio-économique des quartiers.

Analyse des communications avec l'international Dans cette partie, nous nous sommes intéressés à comprendre le comportement des utilisateurs communiquant depuis et vers l'international. Dans un premier temps, on s'est focalisé sur les clusters de pays. Les outils d'analyse exploratoire nous ont permis de nous orienter assez vite vers certains clusters de pays. On a ainsi pu comprendre comment se distribuaient la communauté burkinabaise sur

l'ensemble du pays. Le cluster du Burkina Faso est en effet le plus caractéristique, ce qui n'est pas surprenant puisqu'il s'agit du pays voisin communiquant le plus avec la Côte d'Ivoire. A contrario, la mesure d'intérêt introduite dans le chapitre 4 nous a permis de mettre en évidence des clusters très atypiques malgré un faible nombre de communications.

L'ajout des composantes temporelles et de la nature de la communication dans les analyses ont permis de détecter des clusters de pays avec des comportements très différents. Le Burkina Faso communique avec la Côte d'Ivoire en début de matinée et en début de soirée, ce qui correspond aux horaires raisonnables de la journée, en dehors des horaires habituels de travail. On a donc affaire à une utilisation du réseau non liée à activité professionnelle.

Analyses de mobilité Cette dernière étude sur la Côte d'Ivoire a consisté en une analyse de la mobilité des usagers. On a étudié les connexions aux antennes relais de 50000 utilisateurs choisis aléatoirement et anonymisés, pendant deux semaines. Il en est ressorti d'une première analyse que les usagers du réseau sont très peu mobiles. On obtient donc une structure très diagonale de la matrice de co-occurrences entre utilisateurs et antennes, signe d'une faible mobilité. Nous avons donc isolé les utilisateurs les plus mobiles grâce à un clustering des courbes de répartition du trafic des utilisateurs, sur le rang des antennes les plus utilisées. En procédant ainsi, nous avons supprimé environ 85% des utilisateurs de la base. En ré-appliquant un co-clustering sur les données filtrées, nous nous sommes débarrassés de la corrélation trop forte entre antennes et utilisateurs. En ajoutant les variables jours de la semaine et heure de la journée, nous avons pu comprendre en quoi les utilisateurs d'un groupe étaient similaires. Cela s'est traduit par une zone de résidence commune et des zones de travail plus dispersées dans la ville mais néanmoins bien identifiées.

Perspectives Ces résultats alimentent un projet de recherche mené par la R&D d'Orange, dont l'objectif est d'étudier la valeur des données produites à partir des réseaux et des services opérés par Orange, dans le contexte très spécifique de l'Afrique. Peu de pays africains possèdent en effet des données d'ordre socio-économiques ou démographiques, nécessaires à la conduite de leurs politiques publiques. Orange, en tant qu'opérateur de télécommunications majeur en Afrique, produit un volume important de données numériques et le Challenge D4D (Blondel *et al.*, 2012) lancé par Orange et sa filiale de Cote d'Ivoire en 2012, a montré les nombreuses applications et services imaginés par les chercheurs.

L'analyse réalisée dans ce chapitre montre comment les données des communications mobiles peuvent apporter des enseignements sur les échanges économiques et sociaux au niveau du territoire Ivoirien, sur les relations internationales, et sur les déplacements des populations. Signalons que l'interprétation des cartes produites dans ce chapitre fait l'objet d'une collaboration de recherche en cours entre Orange Labs et l'Université Alassane Ouattara d'Abidjan.

Orange identifie de nombreuses applications de l'analyse de ses données en Afrique, notamment dans le domaine de la planification d'infrastructure urbaine et la dématérialisation des services administratifs (inscriptions des étudiants, déclarations de naissance...), pour répondre aux enjeux de la ville africaine et de la croissance démographique urbaine.

6

Conclusion

Dans cette thèse, nous avons introduit en premier lieu les notions clés du clustering afin de mettre en évidence les difficultés liées à ce type de techniques d'analyse non supervisée. Nous avons ensuite passé en revue différentes méthodes de co-clustering et positionné l'approche MODL par rapport aux concepts introduits dans la littérature. Dans le cas particulier de l'approche MODL, il s'agit de faire un partitionnement conjoint des valeurs prises par les variables – continues ou nominales – descriptives des données.

Différents problèmes de data mining ont été traités dans cette thèse à l'aide de MODL. Nous avons rappelé la formalisation (déjà introduite par Boullé (2011)) du problème de clustering de nœuds dans les graphes. Grâce à des expérimentations sur des données artificielles engendrées suivant deux modèles génératifs différents, nous avons montré la pertinence de l'utilisation de l'approche. Nous avons également vu la diversité des structures pouvant être inférées, en mettant en évidence des performances comparables entre MODL et des approches alternatives, sur des graphes avec des structures favorables aux approches alternatives. Nous avons étendu le problème du partitionnement de graphes aux graphes temporels et proposé une nouvelle formalisation permettant, d'une part, de faire un clustering des nœuds, et d'autre part, de discrétiser le temps en intervalles, dans lesquels le graphe est stationnaire. L'utilisation d'une méthode de triclustering sur ce problème permet de traiter des graphes dont les événements ne sont pas enregistrés à intervalles réguliers dans le temps. Ainsi nous considérons un graphe temporel comme un graphe dont la structure évolue au cours du temps et non comme une succession de graphes statiques, comme c'est le cas dans la plupart des approches alternatives. Des expérimentations sur des graphes engendrés artificiellement ont été menées de manière à montrer la fiabilité de l'approche sur des graphes stationnaires et des graphes avec une évolution temporelle linéaire. Afin d'illustrer la diversité des problèmes pouvant être traités par le triclustering (ou co-clustering en trois dimensions), nous présentons également une application pour le clustering de courbes. Nous avons suivi la même démarche d'introduction de la formalisation et d'expérimentations sur des données artificielles.

L'approche MODL est adaptée au traitement de grands volumes de données comme c'est le cas dans les bases de données textuelles, les comptes-rendus d'appels, etc. Lorsque l'approche de co-clustering est appliquée à des données

volumineuses décrites par des variables prenant un grand nombre de valeurs différentes, les résultats peuvent être très fins et donc difficiles à exploiter. Afin de guider l'utilisateur dans l'interprétation de tels résultats, nous avons développé trois axes d'analyse exploratoire. Le premier consiste à simplifier les résultats à partir de la structure la plus fine jusqu'à la plus grossière, c'est-à-dire la structure avec un seul cluster. Ce post-traitement revient à réaliser une classification hiérarchique ascendante avec la partition optimale en bas de la hiérarchie. Nous avons défini une mesure de dissimilarité entre les clusters afin de déterminer la fusion des clusters la moins coûteuse à chaque étape de la classification hiérarchique. Cette dissimilarité est directement déduite du critère optimisé dans l'approche MODL. Ses propriétés asymptotiques sont analysées et ont permis de faire un lien avec les approches de co-clustering basées sur la théorie de l'information. Le second axe d'analyse exploratoire consiste en un ensemble d'indicateurs permettant de se focaliser sur les motifs intéressants de la structure de co-clustering. Après avoir introduit les notions d'inertie dans les grilles de co-clustering, nous proposons des mesures permettant de déterminer les clusters les plus remarquables et les valeurs les plus représentatives de leur cluster. La détection des clusters remarquables est utile lorsque le nombre de clusters est trop important pour qu'ils soient tous analysés. Quant aux valeurs représentatives, on les utilise pour étiqueter les clusters auxquels elle appartiennent. Enfin, le dernier axe d'analyse exploratoire est la visualisation. Nous avons proposé une visualisation basée sur l'information mutuelle entre les variables pour comprendre la structure de la grille de co-clustering. Cette visualisation nous permet de détecter les co-clusters où il y a des excès et des déficits d'observations. Cette visualisation est adaptée au biclustering mais dès qu'on augmente le nombre de dimensions, la visualisation devient plus complexe. C'est la raison pour laquelle nous avons introduit un critère de contraste, permettant différentes interprétations des résultats lorsqu'on a plus de deux variables descriptives.

Les différentes applications du co-clustering ainsi que les outils d'analyse exploratoire sont illustrés sur un cas pratique en lien avec l'activité de l'opérateur Orange : l'analyse des comptes-rendus d'appels passés en Côtes d'Ivoire entre Décembre 2011 et Janvier 2012. Cette base de données contient environ 500 millions d'observations (ici des appels téléphoniques). Trois types d'études ont été menées : une analyse de comptes-rendus d'appels agrégés par antenne, une analyse du trafic entre les mobiles Ivoiriens et l'international et une étude de mobilité. Toutes les analyses ont été réalisées en appliquant l'approche MODL. Dans de nombreux cas, les résultats se sont montrés trop fins pour être analysés directement. Les outils d'analyse exploratoire introduits dans cette thèse ont alors été utilisés pour permettre une interprétation des résultats, illustrant ainsi leur utilité sur un cas concret. Nous avons vu que, grâce à ces indicateurs, nous pouvions tirer une information utile des résultats. Cette information a été interprétée afin de mieux comprendre le comportement des usagers en termes de téléphonie mobile dans le pays. Ces études sont actuellement utilisées dans le cadre d'un projet industriel mené en partenariat avec des sociologues de

l'Université Alassane Ouattara de Bouaké en Côte d'Ivoire.

Bien que la taille des bases de données étudiées dans la thèse soit importante, leur volumétrie n'est pas suffisante pour en faire des *Big Data*. L'aspect algorithmique n'a pas été traité dans cette thèse. On peut donc se demander quelles sont les limites de l'approche MODL, et des approches de co-clustering en général, en termes de passage à l'échelle. On a vu qu'il était indispensable de disposer d'une quantité de données significative pour mener des études fiables. Mais la collecte et l'analyse de ces données sont coûteuses, ce qui nous amène à nous demander s'il est utile d'augmenter indéfiniment leur volume quand les structures obtenues nécessitent d'être simplifiées pour être interprétées.

Nous avons introduit des outils d'analyse exploratoire qui nous ont permis d'extraire une information exploitable dans l'étude menée sur les comptes-rendus d'appels de Côte d'Ivoire. Des parallèles ont été faits avec des notions de théorie de l'information : les concepts déduits de l'approche MODL ont tous une interprétation asymptotique liée à l'information mutuelle. De nombreuses approches de co-clustering utilisent l'information mutuelle – le plus souvent sous sa forme normalisée – pour mesurer la qualité de la structure obtenue. Certains l'optimisent même directement pour inférer leur structure de co-clustering. On pourrait donc utiliser les outils d'analyse exploratoire dans leur définition asymptotique pour l'analyse des résultats obtenus par des approches alternatives à MODL. L'analyse de corpus de textes est une application courante du co-clustering (Dhillon *et al.*, 2002), pour laquelle l'utilisation de telles techniques d'analyse exploratoire pourrait se montrer efficace, notamment pour filtrer les mots-outils, pour étiqueter les clusters, ajouter des nouveaux textes, etc.

Dans cette thèse, on s'est limité à des études simplement descriptives. On pourrait envisager d'utiliser des résultats de l'analyse exploratoire pour l'enrichissement des bases de données, utile pour des analyses supervisées comme le scoring par exemple. Dans le cas de la Côte d'Ivoire, on a vu que la valeur des indicateurs construits sur les résultats de co-clustering ont une interprétation sociologique. Cette interprétation a été validée par des experts du domaine. Il est ainsi envisageable de réaliser un co-clustering sur des bases de données, d'attribuer un score aux profils, de détecter les individus les plus représentatifs de leur profil et d'utiliser ces informations afin d'enrichir les bases de données, ce qui permettrait d'améliorer la connaissance des usages, de développer l'activité ou encore de d'améliorer les infrastructures urbaines, notamment liées à la mobilité.

L'approche MODL permet de réaliser un co-clustering en d dimensions. Le problème est que le nombre d'observations nécessaires pour peupler une grille de co-clustering croît exponentiellement avec le nombre de dimensions. Dans certains problèmes, les données sont décrites par de nombreuses variables mais le faible nombre d'observations ne permet pas à MODL d'inférer une structure. Ce point a été illustré dans le chapitre 5 et a été l'objet d'une discussion quant à l'utilisation directe du triclustering dans l'analyse temporelle du graphe d'appels. On peut imaginer plusieurs axes de réflexions afin de résoudre ce problème.

Une solution serait d'utiliser MODL dans une optique de co-clustering de variables et d'observations – ce qui correspond à l'application la plus courante du co-clustering – afin de simplifier les données en amont de l'application d'un co-clustering entre variables. Cependant, ce pré-traitement se limite aux cas de variables continues. On pourrait également imaginer construire une hiérarchie de co-clusterings. Il s'agit de réaliser des co-clusterings sur plusieurs sous-ensembles de variables distincts et d'utiliser les co-partitions obtenues pour créer de nouvelles variables sur lesquelles on applique un nouveau co-clustering. L'opération est réitérée jusqu'à ce qu'il n'y ait qu'un seul co-clustering de toutes les variables.

Ce type d'évolutions permettrait de structurer l'ensemble des variables et de restituer à l'utilisateur des résultats plus simples à analyser et plus informatifs, ce qui est l'objectif premier du clustering.

Bibliographie

M. Abramowitz et I.A Stegun : Handbook of mathematical functions with formulas, graph, and mathematical tables. *Applied Mathematics Series*, 55:1046, 1965.

E.M. Airoldi, D.M. Blei, S.E. Fienberg et E.P. Xing : Mixed Membership Stochastic Blockmodels. *JMLR*, 9:1981–2014, 2008.

H. Akaike : A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

C. Ambroise, G. Grasseau, M. Hoebeke et Latouche P. : The mixer package. 2010.

P. Arabie, S. Schleutermann, J. Daws et L. Hubert : Marketing applications of sequencing and partitioning of nonsymmetric and/or two-mode matrices. pages 215–224, 1988.

B.B. Baker et E.T. Copson : *The mathematical theory of Huygens' principle*. Clarendon Press Oxford, 1950.

A. Banerjee, S. Merugu, I.S. Dhillon et J. Ghosh : Clustering with Bregman divergences. *In Journal of Machine Learning Research*, 2004.

V. Batagelj, P. Doreian et A. Ferligoj : An optimizational approach to regular equivalence. *Social Networks*, 14:121–135, 1992.

J.C. Bezdek et R.J. Hathaway : VAT : a tool for visual assessment of (cluster) tendency. *In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, volume 3, pages 2225–2230. IEEE, 2002.

P.J. Bickel et A. Chen : A nonparametric view of network models and Newman Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50), 2009.

C. Blake et C.J. Merz : {UCI} repository of machine learning databases. 1998.

D.M. Blei et M.I. Jordan : Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

D.M Blei, A.Y. Ng et M.I. Jordan : Latent Dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

V.D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda et C. Ziemlicki : Data for development : the d4d challenge on mobile phone data. 2012.

V.D Blondel, J-L. Guillaume, R. Lambiotte et E. Lefebvre : Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008(10), 2008.

V.D. Blondel, G. Krings et I. Thomas : Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. 2010.

S. P. Borgatti : A comment on Doreian's regular equivalence in symmetric structures. *Social Networks*, 10, 1988.

S. Boriah, V. Chandola et V. Kumar : Similarity measures for categorical data : A comparative evaluation. *In SDM*, pages 243–254, 2008.

M. Boullé : *Recherche d'une représentation des données efficace pour la fouille de grandes bases de données.* Thèse de doctorat, École Nationale des Télécommunications, 2007.

M. Boullé : Bivariate data grid models for supervised learning. Rapport technique, Orange Labs, 2008.

M. Boullé : Estimation de la densité d'arcs dans les graphes de grande taille : une alternative à la détection de clusters. *In Extraction et gestion des connaissances (EGC'2011)*, pages 353–364, 2011.

M. Boullé : Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition*, 45(12):4389–4401, 2012.

M. Boullé : Sélection bayésienne de modèles avec prior dépendant des données. *In Extraction et gestion des connaissances (EGC'2012)*, pages 29–34, 2012.

M. Boullé, R. Guigourès et F. Rossi : Nonparametric hierarchical clustering for functional data. *In Advances in Knowledge Discovery and Management.* 2013.

U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski et D. Wagner : *On modularity-np-completeness and beyond.* 2006.

R.L. Breiger, S.A. Boorman et P. Arabie : An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3), 1975.

F. Chamroukhi, A. Samé, G. Govaert et P. Aknin : A hidden process regression model for functional data description. application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, March 2010.

M. Charrad et M. Ben Ahmed : Simultaneous clustering : A survey. *In Pattern Recognition and Machine Intelligence*, pages 370–375. Springer, 2011.

Y. Cheng et G.M. Church : Biclustering of expression data. *In Proceedings of the eighth international conference on intelligent systems for molecular biology*, volume 8, pages 93–103, 2000.

C.H. Coombs, R.M. Dawes et A. Tversky : *Mathematical psychology : An elementary introduction.* Prentice-Hall Englewood Cliffs, NJ, 1970.

T.M. Cover et J.A. Thomas : *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

M.n Dash, K. Choi, P. Scheuermann et H. Liu : Feature selection for clustering-a filter solution. *In ICDM 2003*, pages 115–122. IEEE, 2002.

T. De Bie : An information theoretic framework for data mining. *In KDD*, pages 564–572, 2011.

I.S. Dhillon : Co-clustering documents and words using bipartite spectral graph partitioning. *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

I.S. Dhillon, S. Mallela et R. Kumar : Enhanced word clustering for hierarchical text classification. *In KDD '02 : Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 191–200. ACM Press, 2002.

I.S. Dhillon, S. Mallela et D.S. Modha : Information-theoretic co-clustering. *In KDD '03*, pages 89–98, 2003.

P.J. Diggle : Statistical analysis of spatial point patterns. 1983.

P. Doreian, V. Batagelj et A. Ferligoj : Generalized blockmodeling of two-mode network data, 2004.

R.O. Duda, P.E. Hart et D.G. Stork : Unsupervised learning and clustering. *Pattern classification*, page 571, 2001.

J.G. Dy et C.E. Brodley : Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

T. Eckes et P. Orlik : An error variance approach to two-mode hierarchical clustering. *Journal of Classification*, 10(1):51–74, 1993.

S. Fortunato : Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

C. Fraley et A.E. Raftery : How many clusters ? Which clustering method ? Answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

M. Friendly : Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

A.E. Gelfand et A.F.M. Smith : Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410): 398–409, 1990.

L. Geng et H.J. Hamilton : Interestingness measures for data mining : A survey. *ACM comput. Surv.*, 38(3), 2006.

M. Girvan et M.E.J. Newman : Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

R.Y. Gnabéli : La production d'une identité autochtone en côte d'Ivoire. *Journal des anthropologues. Association française des anthropologues*, (114-115):247–275, 2008.

A. Goldenberg, A. X. Zheng, S.E. Fienberg et E. M. Airoldi : A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

G. Govaert : Algorithme de classification d'un tableau de contingence. *In First international symposium on data analysis and informatics*, pages 487–500, Versailles, 1977. INRIA.

G. Govaert : Simultaneous clustering of rows and columns. *Control and Cybernetics*, 24(4):437–458, 1995.

G. Govaert et M. Nadif : Clustering with block mixture models. *Pattern Recognition*, 36:463–473, 2003.

G. Govaert et M. Nadif : *Co-Clustering*. John Wiley & Sons, 2013.

P.D. Grünwald : *The Minimum Description Length Principle*. Mit Press, 2007.

R. Guigourès et M. Boullé : Segmentation of towns using call detail records. *In NetMob Workshop at IEEE SocialCom*, 2011.

R. Guigourès, M. Boullé et F. Rossi : A triclustering approach for time evolving graphs. *In ICDM Workshops*, pages 115–122, 2012.

R. Guigourès, M. Boullé et F. Rossi : Étude des corrélations spatio-temporelles des appels mobiles en France. *In Extraction et gestion des connaissances (EGC 2013)*, pages 437–448, 2013.

I. Guyon et A. Elisseeff : An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

P. Hansen et N. Mladenovic : Variable neighborhood search : Principles and applications. *European Journal of Operational Research*, 130(3):449–467, 2001.

J.A. Hartigan : Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

T. Hastie, R. Tibshirani et J. Friedman : *The Elements of Statistical Learning*. Springer, 2009.

G. Hébrail, B. Hugueney, Y. Lechevallier et F. Rossi : Exploratory Analysis of Functional Data via Clustering and Optimal Segmentation. *Neurocomputing*, 73(7-9):1125–1141, 2010.

J.L. Hintze et R.D. Nelson : Violin plots : a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

P. W. Holland, K. Laskey et S. Leinhardt : Stochastic blockmodels : First steps. *Social Networks*, 5(2):109–137, 1983.

J. Hopcroft, O. Khan, B. Kulis et B. Selman : Tracking evolving communities in large linked networks. *PNAS*, 101:5249–5253, 2004.

L. Hubert et P. Arabie : Comparing partitions. *Journal of classification*, 2 (1):193–218, 1985.

A.K. Jain : Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

A.K. Jain et R.C. Dubes : *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola et L.K. Saul : An introduction to variational methods for graphical models. *Machine learning*, 37 (2):183–233, 1999.

B. Karrer et M.E.J. Newman : Stochastic blockmodels and community structure in networks. *CoRR*, abs/1008.3926, 2010.

C. Kemp et J.B. Tenenbaum : Learning systems of concepts with an infinite relational model. *In AAAI'06*, 2006.

Y. Kluger, R. Basri, J.T. Chang et M. Gerstein : Spectral biclustering of microarray data : coclustering genes and conditions. *Genome research*, 13 (4):703–716, 2003.

S. Kullback et R. A. Leibler : On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

P. Latouche, E. Birmelé et C. Ambroise : Bayesian methods for graph clustering. *In Advances in Data Analysis, Data Handling and Business Intelligence*, pages 229–239. Springer, 2010.

J. Leskovec, J. Kleinberg et C. Faloutsos : Graphs over time : densification laws, shrinking diameters and possible explanations. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

Y.H. Li et A.K. Jain : Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

J. Lin : Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37:145–151, 1991.

H. Liu et L. Yu : Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, 2005.

F. Lorrain et H.C. White : Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1(49-80), 1971.

S.C. Madeira et A.L. Olivieira : Biclustering algorithms for biological data analysis : a survey. *Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

G.W. Milligan et M.C. Cooper : An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

B. Mirkin : *Mathematical classification and clustering*. Kluwer Academic Press, 1996.

B. Mirkin, P. Arabie et L. Hubert : Additive two-mode clustering : the error-variance approach revisited. *Journal of Classification*, 12(2):243–263, 1995.

S. F. Nadel : *The Theory of Social Structure*. Cohen & West, 1957.

M. Nadif et G. Govaert : Model-based co-clustering for continuous data. *In ICMLA*, pages 175–180, 2010.

R.M. Neal : Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational AND Graphical Statistics*, 9(2):249–265, 2000.

A.Y. Ng, M.I. Jordan et Y. Weiss : On spectral clustering : Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

X.L. Nguyen et A.E Gelfand : The Dirichlet labeling process for clustering functional data. *Sinica Statistica*, 21(3):1249–1289, 2011.

K. Nowicki et T. Snijders : Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96:1077–1087, 2001.

G. Palla, A-L Barabási et T. Vicsek : Quantifying social group evolution. *Nature*, 446, 2007.

G. Palla, I. Derenyi, I. Farkas et T. Vicsek : Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

J. Pitman : *Combinatorial stochastic processes*, volume 1875 de *Lecture Notes in Mathematics*. Springer-Verlag, 2006.

A. Pothen, H.D. Simon et K-P. Liou : Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990.

H. Ralambondrainy : A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16(11):1147–1157, 1995.

J.O. Ramsay et B.W. Silverman : *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005.

W.M. Rand : Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

J. Reichardt et D. R. White : Role models for complex networks. *The European Physical Journal B*, 60, 2007.

J. Rissanen : Modeling by shortest data description. volume 14, pages 465–471, 1978.

F. Rossi et N. Villa-Vialaneix : Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets. *Journal de la Société Française de Statistique*, 152(3):34–65, 2012.

P.J. Rousseeuw : Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

S.E. Schaeffer : Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

H. Shan et A. Banerjee : Bayesian co-clustering. *In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 530–539. IEEE, 2008.

C. E. Shannon : A mathematical theory of communication. *Bell system tech. journal*, 27, 1948.

A. Silberschatz et A. Tuzhilin : On subjective measure of interestingness in knowledge discovery. *In KDD*, pages 275–281, 1995.

N. Slonim et N. Tishby : Document clustering using word clusters via the information bottleneck method. *In In ACM SIGIR 2000*, pages 208–215. ACM Press, 2000.

A.M. Steane : Error correcting codes in quantum theory. *Physical Review Letters*, 77(5):793, 1996.

A. Strehl et J. Ghosh : Cluster ensembles – a knowledge reuse framework for combining multiple partition. *JMLR*, 3:583–617, 2003.

J. Sun, C. Faloutsos, S. Papadimitriou et P.S. Yu : Graphscope : parameter-free mining of large time-evolving graphs. KDD '07, pages 687–696, 2007.

Y. W. Teh : Dirichlet processes. *In Encyclopedia of Machine Learning*. Springer, 2010.

R. Tibshirani, G. Walther et T. Hastie : Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society - Series B : Statistical Methodology*, 63(2):411–423, 2001.

I. Van Mechelen, H.-H. Bock et P. De Boeck : Two-mode clustering methods : a structured overview. *Statistical methods in medical research*, 13 (5):363–394, 2004.

J.E. Vogt, S. Prabhakaran, T.J. Fuchs et V. Roth : The translation-invariant Wishart-Dirichlet process for clustering distance data, 2010.

C.S. Wallace et D.M. Boulton : An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

H.M. Wallach, S.T. Jensen, L.D. et K.A. Heller : An alternative prior process for nonparametric bayesian clustering. *In AISTATS*, pages 892–899, 2010.

S. Wasserman et K. Faust : *Social Network Analysis : Methods and Applications*. Structural analysis in the social sciences. Cambridge Univ. Press, 1994.

D. R. White et K. P. Reitz : Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2), 1983.

H.C. White, S. Boorman et R. Breiger : Social structure from multiple networks : I. blockmodels of roles and positions. *Am. J. of Sociology*, 81 (4):730–80, 1976.

R.D. Wilson et T.R. Martinez : Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

E. P. Xing, W. Fu et L. Song : A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 4(2):535–566, 2010.

L. Zhao et M.J. Zaki : Tricluster : An effective algorithm for mining coherent clusters in 3d microarray data. *In SIGMOD Conference*, pages 694–705, 2005.